UNDERSTANDING AND MODELING USER BEHAVIOR IN SOCIAL QUESTION AND ANSWERING

A Dissertation in
Information Sciences and Technology

by

Zhe Liu

© 2015 Zhe Liu

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2015
The dissertation of Zhe Liu was reviewed and approved* by the following:

Bernard J. (Jim) Jansen
Professor, Information Sciences and Technology
Dissertation Advisor
Chair of Committee

John Yen
Professor, College of Information Sciences and Technology

Heng Xu
Associate Professor, College of Information Sciences and Technology

James L. Rosenberger
Professor, Department of Statistics

Carleen Maitland
Interim Associate Dean for Undergraduate and Graduate Studies
Associate Professor, College of Information Sciences and Technology

*Signatures are on file in the Graduate School
ABSTRACT

In this research, I focus on a specific type of information seeking on social networking sites (SNS), called social question and answering (social Q&A), in which people ask natural language questions to their networks via status updates. Compared with the traditional information seeking methods (e.g. search engines, online databases, etc.) and community question answering services (e.g. Yahoo! Answers, Quora, Answers.com, StackOverflow, etc.), social Q&A is considered to provide more trustworthy, contextualized, and interactive responses. Although there is an extensive amount of research and literature available on the topic of information seeking via the Internet, due to the distinct nature of social Q&A, still relatively little is known about how users behave in the process of information seeking and sharing under social context. To fill in this gap, in this study I examine three aspects of social Q&A.

First, in order to better understand the information needs of the questioners in a microblogging environment, I develop a taxonomy of question types in social Q&A, called ASK. This taxonomy allows us to differentiate questions according to their underlying intents, and then direct them to the most appropriate respondents for help. To apply the taxonomy to practical problems, I also present the implementation of an automatic classification method to categorize the social Q&A questions according to the ASK taxonomy.

Second, noticing the low response rate in social Q&A, I examine a set of factors that would affect the response probability of a question. Specifically, I limit the scope to extrinsic factors only, given the existence of literature focusing on factors from both social and cognitive perspectives in knowledge sharing. For the analysis, I collect over 20 thousand real-world questions posted on Weibo, which is the largest microblogging site in China.

Third, to validate the effectiveness of question routing systems in social Q&A, as well as to characterize the voluntary knowledge sharing behavior among individuals under social context,
I analyze a collection of questions and answers posted during a 10-month period on Weiwen, a Chinese question routing services based on microblogging sites. I explore the patterns demonstrated by the knowledge contributors from three different perspectives: user behavior, user interest, and user connectedness. I also propose a predictive model on active contributors in social Q&A environment based on a number of non-Q&A features.

I believe this thesis is of interest as it very well addresses the research gaps concerning the understanding of information needs in social Q&A, and the effectiveness of SNS in handling question-answering. Findings from this research would be of practical value as well, as it could fulfill the need for technologies capable of performing question routing tasks to help people find what they need.
# TABLE OF CONTENTS

List of Figures .................................................................................................................. vii

List of Tables ................................................................................................................... ix

Acknowledgements .......................................................................................................... x

Chapter 1 Introduction ..................................................................................................... 1
  Research Motivations ................................................................................................. 3
  Research Objectives ................................................................................................. 5
  Thesis Organization ................................................................................................. 7

Chapter 2 Literature Review .......................................................................................... 9
  Community Question and Answering ...................................................................... 9
  Social Question and Answering ............................................................................... 11
  Questions Asked in Social Q&A ............................................................................... 12
  Questions Answered in Social Q&A ........................................................................ 13
  Social Q&A Systems .............................................................................................. 14

Chapter 3 ASK: A Taxonomy of Information Seeking Posts in Social Question and Answering ......................................................... 16
  Background .............................................................................................................. 16
  Taxonomy of social Q&A ....................................................................................... 16
  Automatic Question Classification ...................................................................... 17
  Research Objectives ............................................................................................. 20
  Methods ....................................................................................................................... 21
  Data Collection ....................................................................................................... 21
  Taxonomy Creation .................................................................................................. 22
  Question Classification Using ASK Taxonomy ...................................................... 25
  Results ....................................................................................................................... 29
  Classification Results using Lexical Features ......................................................... 29
  Classification Results using Topical Features ....................................................... 31
  Classification Results using Contextual Features ................................................ 34
  Classification Results using Syntactic Features .................................................... 36
  Ensemble Classification Results using Features from the Question’s Perspective .. 37
  Classification Results using Features from the Answer’s Perspective .................... 38
  Discussion and Design Implications ...................................................................... 42
  Conclusion ................................................................................................................ 44

Chapter 4 Predicting the Response Probability in Social Question and Answering on Sina Weibo .......................................................... 46
  Background .............................................................................................................. 47
<table>
<thead>
<tr>
<th>Chapter 5 Identifying the Willingness to Response: Characterizing Knowledge Contribution of Individuals in Social Q&amp;A</th>
<th>72</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>73</td>
</tr>
<tr>
<td>Research Objectives</td>
<td>74</td>
</tr>
<tr>
<td>Weiwen and Data Set</td>
<td>75</td>
</tr>
<tr>
<td>Results Characterizing Knowledge Sharing Behavior on Weiwen</td>
<td>77</td>
</tr>
<tr>
<td>Aggregate Analysis of Question and Answering</td>
<td>77</td>
</tr>
<tr>
<td>Topical Distribution</td>
<td>79</td>
</tr>
<tr>
<td>Network Analysis</td>
<td>85</td>
</tr>
<tr>
<td>Predicting Super Contributors using Non-Q&amp;A Features</td>
<td>88</td>
</tr>
<tr>
<td>Data Set</td>
<td>89</td>
</tr>
<tr>
<td>Feature Engineering</td>
<td>90</td>
</tr>
<tr>
<td>Classification Algorithms and Evaluation Metrics</td>
<td>92</td>
</tr>
<tr>
<td>Classification Results</td>
<td>93</td>
</tr>
<tr>
<td>Discussion and Implications</td>
<td>97</td>
</tr>
<tr>
<td>Conclusion</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 6 Conclusion</th>
<th>102</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thesis Conclusion</td>
<td>102</td>
</tr>
<tr>
<td>Understanding user intent in Social Q&amp;A</td>
<td>102</td>
</tr>
<tr>
<td>Analyzing Question Response Probability in Social Q&amp;A</td>
<td>104</td>
</tr>
<tr>
<td>Characterizing Knowledge Contribution in Social Q&amp;A</td>
<td>105</td>
</tr>
<tr>
<td>Research Contributions</td>
<td>106</td>
</tr>
<tr>
<td>Future Research</td>
<td>107</td>
</tr>
<tr>
<td>Running Experiment based on the Current Findings</td>
<td>108</td>
</tr>
<tr>
<td>Improvement on the Question Routing Models</td>
<td>108</td>
</tr>
<tr>
<td>Development of a Mobile-based Social Q&amp;A Platform</td>
<td>109</td>
</tr>
<tr>
<td>Closing Remarks</td>
<td>110</td>
</tr>
</tbody>
</table>

| References                                                 | 111|
LIST OF FIGURES

Figure 1-1. Examples of social Q&A questions on Twitter.......................................................... 3
Figure 1-2. Areas of research in social Q&A (Shah et al., 2009)................................................. 7
Figure 3-1. Lexical feature selection using information gain....................................................... 30
Figure 3-2. Distribution of question length on word and character levels across ASK types. ................................................................................................................................. 40
Figure 3-3. Distribution of question response time in log (seconds) across ASK types. ........41
Figure 3-4. Hybrid social Q&A system leveraging on ASK taxonomy for question routing. ......................................................................................................................................... 44
Figure 4-1. Layout of Weibo interface highlighting the replying aspect of threaded answers to posted questions, with various aspect of the interface annotated. ........... 50
Figure 4-2. Distribution of responses to questions posted on Weibo illustrating a long tail distribution. ........................................................................................................................................... 59
Figure 4-3. Cumulative distribution function of two question uniqueness measurements..... 64
Figure 4-4. Distribution of topical categories of questions and their response probability..... 65
Figure 4-5. Distribution of posting time period of questions and their response probability. ................................................................................................................................. 66
Figure 4-6. Box plots depicting the distribution for responded and non-responded social Q&A questions based on retweeting behaviors and past social activities. .......... 67
Figure 5-1. Illustration of the question routing procedure of Weiwen. .................................. 76
Figure 5-2. Screenshot of Weiwen with the major sections of Question Routed, Answers Received, and Non-responsers highlight, along with the Question Category. .......... 77
Figure 5-3. Distributions of the number of questions posted and the number of answers provided on Weiwen in the 10-month data collection period...................................... 79
Figure 5-4. Topical distribution of questions asked on Weiwen by percentage of questions posted by topic, number of answers received by topic, and word length of both questions and answers.......................................................... 80
Figure 5-5. Distribution of number of topics answered and questions answered, showing high correlation between responsiveness and interest diversity. ......................... 82
Figure 5-6. The NPMI values between topic pairs, with associations displayed in seven level scale from (-0.3, -0.4) to (0.9, 1.0), lowest to highest association......................... 84
Figure 5-7. The Bow Tie Structure (Broder et al., 2000) .......................................................... 86

Figure 5-8. Difference in log scaled number of followers and reciprocal relations for active contributors and non-active contributors......................................................... 95

Figure 5-9. Difference in percentage of mention, URL sharing, and retweet posts for active contributors and non-active contributors.......................................................... 96

Figure 5-10. Difference in percentage of posts received like, retweet, or any type of interaction for active contributors and non-active contributors................................. 97
LIST OF TABLES

Table 3-1. Annotation criteria for ASK question types and non-information seeking questions
...................................................................................................................... 24

Table 3-2. Performance of the lexical-based classifier ........................................ 30

Table 3-3. Top 10 n-gram features of each ASK question type along with the informational gain
.......................................................................................................................... 31

Table 3-4. Fraction of ASK questions within each topical category .................. 32

Table 3-5. Performance of the topical-based classifier ........................................ 34

Table 3-6. Distribution of ASK questions with and without location or temporal mentions
.................................................................................................................................. 34

Table 3-7. Top 5 question topics by percentage of location mentions ................ 35

Table 3-8. Top 5 question topics by percentage of temporal mentions ............... 35

Table 3-9. Performance of the general contextual-based classifier .................... 36

Table 3-10. Performance of the syntactic-based classifier ................................... 36

Table 3-11. Question length across ASK question types ..................................... 37

Table 3-12. Performance of the ensemble-based classifier ................................ 38

Table 3-13. Descriptive Statistics of answered questions across ASK types ....... 38

Table 3-14. Performance of the answer’s perspective-classifier ......................... 38

Table 4-1. Results of logistic regression with coefficient, odds ratio, and p-value .... 42

Table 4-2. Top 10 emoticons adopted in responded questions ......................... 62

Table 4-3. Cross-tabular distribution of @mention and responded type ............. 62

Table 4-4. Classification results with the logistic regression models using user-based, question-based, social activity-based, and all features ................................................. 69

Table 5-1. Connectedness statistics for questions within 13 topical categories in Weiwen .... 87

Table 5-2. Classification results using Naïve Bayes, SMO, and Logistic Regression .... 93

Table 5-3. Results of Logistic regression with coefficients and p-values of each feature .... 94
ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Jim Jansen for his supportive mentorship and sagacious guidance. Thank you for always being there and helping me through my PhD journey. It would not be possible to complete this dissertation without your invaluable insights and suggestions.

I would also like to thank Dr. John Yen, Dr. Heng Xu, and Dr. James L. Rosenberger. Thank you for serving on my dissertation committee and for your insightful comments about my work.

I am grateful to Dr. Ingmar Weber and Yingxin Pan for being great mentors during my internships at Qatar Computing Research Institute and IBM China Research Laboratory. Thank you for providing me the intellectually inspiring environment within which I could work on numerous projects.

My sincere appreciation also goes to my friends for your care and companion. Without you my PhD life would have been far less exciting.

Last but most importantly, I reserve the greatest acknowledgement for my beloved family - my parents and grandparents, to whom I dedicate this dissertation. Thank you for your understanding, constant encouragement and endless love across continents.
Chapter 1

Introduction

Information seeking is the process or activity of attempting to obtain information in both human and technological contexts (Shih, Chen, Chu, & Chen, 2012). Long before the invention of computers, the most prevailing information seeking option was the use of other people as information sources. People tend to enjoy such a quick way of information seeking due to several reasons, including the readily source accessibility (Case, 2012), as well as highly personalized results. However, studies indicated that the use of human as information sources is more of a least effort option in the search for information, therefore, sometimes those human sources may not be the best choice available (Childers & Post, 1975). The introduction of computers and Internet technologies has completely changed people lives in the recent decades.

To comply with the trend of computerization, digital information seeking tools, such as search engines, online catalogs, question-and-answering sites, etc., have been develop to accommodate people’s more complex information needs. All of these tools have made the Internet a powerful source for information acquisition, and therefore led to the shift in people’s information seeking habits from human-based search to machine-based search. Although the machine-based information seeking approaches today are faster and more efficient than ever, it still suffers from the problem of being ambiguous and lacking in context.

Given that both pure human-based and machine-based searching techniques have some weaknesses, recent studies have suggested a new information seeking strategy, called social question and answering (social Q&A). The huge rise in the popularity of social networking sites (SNS), such as Facebook, Twitter, and Google+ etc., in the past decade has made them preferable platforms for people to catch up on news and real-time events that happened to their social ties.
With hundreds of millions of messages transmitted every day, SNS not only provide people with the opportunities to connect and share with the world, but also become rich data sources for information seeking. Defined as the process of “information seeking by asking natural language questions to other users in a social network” (Shah, Oh, & Oh, 2008), social Q&A lies between the boundaries of human and machine-powered information seeking models.

On the one hand, social Q&A frames information seeking as a more natural activity as compared to the machine-based approach. Information seekers using search engines often need to go through the results and select documents relevant to their information needs (Shah et al., 2008). However, social Q&A allows individuals to simply post natural language questions to their network via status updates. In this way, social Q&A generates more personalized and interactive answers, specially tailored to the questioner’s individual contexts. In addition to that, the profile-based user accounts of SNS makes the platform non-anonymous and therefore enables the information seekers to better trust the posted contents (Svensson, 2011). On the other hand, social Q&A also surpasses the temporal and geographical limits of the pure human-based approach by allowing the questioners and the answerers to exchange information asynchronously online. Due to such asynchronous nature of social Q&A, all potential answerers can easily participate in the Q&A process or view the content at their own pace. Besides, leveraging on the wide reach of the Internet, social Q&A changed the question-answerer relationship in traditional human-based search from on-to-one to one-to-many. In this way, the efficiency of question dissemination can be greatly improved. Some of the social Q&A examples are shown in Figure 1-1, include: Examples of social Q&A include: Anyone knows how to fix blinking monitor?; Can anyone recommend any good places to go for afternoon tea in central London?; #healthadvice Twitter I need help - how can I kick a cold/flu illness quickly?
Like other Web 2.0 services, social Q&A takes the advantage of the “wisdom of crowds” (Surowiecki, 2004), and facilitate the collaborative information seeking and sharing over the SNS. Shah, Oh, and Oh (2009) suggested that along with people’s adoption of social Q&A services, information seeking habits have also been changed to more customized information than that obtained by keyword-based queries through a traditional search engine. It is therefore important to provide a comprehensive understanding of how users behave in social Q&A process, especially from a design perspective, which has until now essentially been overlooked in many of the past studies.

With the growing evidence on the effectiveness of social Q&A in practice, it has recently received significant attention from the industry on evaluating its potential value in future search. According to Jansen, Liu, Weaver, Campbell, and Gregg (2011), Google and other search engine companies already realized the potential of real time postings and are experimenting with methods to archive this social media content. Yet, in the same time, the social networking giant Facebook have rolled out a beta feature called Facebook Questions. By taking advantage of its
own resources, Facebook Questions offers its users a quick way to survey their friends in the format of multiple-choice questions and thus enabling them to pull recommendations from those like-minded ones (Hicks, 2010). However, Facebook discontinued this question feature after only a year on the market due to its failure to gain traction among its users. This failure leads me to rethink whether the current design of SNS or social Q&A tools or service can adequately meet the needs of individuals who ask or answer questions on it, and how we can better understand individual’s social information seeking and sharing habits while analyzing through large scale social Q&A activities.

To be more specific, I identify two major gaps in the current design of social Q&A systems, as well as in existing literatures.

**Research gap 1: Lack of Evaluation on the Current Social Q&A Services**

Although social Q&A is now viewed as the next major revolution that will change the entire landscape of the Internet, very few of the existing studies have examined the performance of SNS with Q&A intents. Among the limited number of studies, Shah and Kitzie (2012) found that, since the SNS is originally designed as a platform for information dissemination rather than a tool for information seeking, there is no guarantee that the question posted on it will be answered. Paul, Hong, and Chi (2011a) in their study pointed out that in general only 23% of questions posted on Twitter receive a response, which is only 1/3 of the response rate of community Q&A sites. Together with the failure of Facebook Questions, all these findings lead us to propose research questions evaluate the effectiveness of SNS in handling questioning and answering.

**Research gap 2: Lack of Design Implication on Social Q&A Systems**
Another research gap that I address in this thesis is the lack of design implications on the effective systems or tools that can facilitate individual’s social Q&A process. Designing effective Q&A systems has been a challenging problem in general. Hsieh (2009) once pointed out the inefficiency existed in the current design of community Q&A sites, indicating that services like Yahoo!Answers were not well designed to accommodate differences in individuals’ needs and constraints as they demonstrated in information seeking process. Although there are a number of studies conducted on developing tools or models to facilitate the Q&A process on community Q&A sites, like Yahoo!Answers and AnswerBag etc (Baichuan Li, Si, Lyu, King, & Chang, 2011; X. Liu, Croft, & Koll, 2005; Pomerantz, Nicholson, & Lankes, 2003), only a few tools or models have been proposed to help the users to find information they need on SNS (Hecht, Teevan, Morris, & Liebling, 2012). Considering the differences in question asking and answering on a community Q&A site versus on a SNS (Morris, Teevan, & Panovich, 2010b), I believe that there is a need for design implications on systems and tools capable of performing question routing or answerer recommending tasks to help people find what they need on SNS.

**Research Objectives**

With these concerns in mind, I propose the following six research objectives from three different perspectives that attempt to fill in the relevant defects in the current design of social Q&A systems, as well as the gaps in existing literatures.

**Research Objective 1 (a):** Develop a taxonomy of questions proposed in social Q&A that could be used to assist in selecting the most appropriate answering strategies.

**Research Objective 1 (b):** Implement the proposed taxonomy by automatically classifying questions into ASK types and measure the effectiveness of the classification.
Research Objective 2 (a): Identify the extrinsic factors that are likely to influence the question response rate in social Q&A.

Research Objective 2 (b): Build a model to predict the question response probability using the extrinsic features.

Research Objective 3 (a): Explore the question answering patterns of individuals when they are exposed to questions asked by strangers via question routing.

Research Objective 3 (b): Identify individuals with the desire to help others in social Q&A by using their non-Q&A characteristics.

Aiming at the above research questions, I conduct studies from three perspectives as suggested by Shah et al. (2009), namely: user, information, technology. Figure 1-2 depicts the research framework that I adopt in this thesis. To be more specific, first, from the user’s aspect, I propose Research Objective 1 to understand individual’s information needs and intents in social Q&A; second, from the content aspects, Research Objective 2 is introduced aiming at understanding the cost and quality of questions in attracting responses from a wide range of audience. At last, I present Research Objective 3 from a technology perspective to evaluate the performance of an existing question routing framework within social context.
Thesis Organization

The remainder of the thesis is organized as follows:

- Chapter 2 covers a comprehensive review of prior studies from three major fields, including community question and answering (community Q&A), social networking sites, and social Q&A.

- Chapter 3, Chapter 4, and Chapter 5 introduce my studies of social Q&A in real world settings, with completed experiments implementation, results and implications. I summarize the contributions and implications of the thesis in Chapter 6.
• In Chapter 3 – “ASK: A Taxonomy of Information Seeking Posts in Social Question and Answering”, I describe my first study in response to the first research objective, exploring the underlying intents behind questions asked in social Q&A.

• In Chapter 4 – “Predicting the Response Rate in Social Question and Answering on Sina Weibo”, I propose another work focusing on my research objective 2, in which I evaluate extrinsic factors that may influence the user’s answering behavior viewed from the quantity perspective.

• In Chapter 5 – “Identifying the Willingness to Response: Characterizing Knowledge Contribution of Individuals in Social Q&A”, meets my third research objective, in which I present a study analyzing questions and answers posted within a question routing framework built based on SNS.

• Chapter 6 summarizes the contributions and broader implications of the thesis.
Chapter 2

Literature Review

Community Question and Answering

According to Harper, Raban, Rafaeli, and Konstan (2008), a community question and answering (community Q&A) site is web site designed purposefully to allow people to ask and answer questions on a broad range of topics. Compared with traditional online reference services, community Q&A sites, such as Yahoo!Answers and Answer Bag tend to leverage more on human resources to answer people’s questions. In other words, community Q&A service allows users to seek for information in a natural way to expressing their information needs as questions rather than keywords, and the answers received as text rather than lists of documents (Shah et al., 2008). Questioners of online Q&A sites can benefit from the “wisdom of crowds” through asking their questions to the public (Surowiecki, 2004).

Community Q&A sites have become the focus of intensive studies in the recent years for its role in providing more personalized information seeking answers. In their study, Harper, Weinberg, Logie, and Konstan (2010) analyzed real-world questions from three popular Q&A sites: Yahoo Answers, Answerbag, and Ask Metafilter. They found that advise-seeking or conversational questions tend to receive more responses than factual questions on those online Q&A sites. Gardelli and Weber (2012) in another study used toolbar data to understand why people submit new questions to Q&A sites. By comparing questioners’ general web usage with their question asking behavior, the authors concluded that users who first issue a question-related query are more likely to issue informational questions, rather than conversational ones on Q&A sites, and such questions are less likely to attract an answer. Bian, Liu, Agichtein, and Zha (2008) found that the majority of the answers received through online Q&A sites reflect personal, often
unsubstantiated opinions. However, questioners tend to prefer personalized answers to their questions.

Given that the quality of user-generated content on online Q&A sites varies across different answers and sites (Harper et al., 2008; Shachaf, 2009), quality assessment is of great importance in providing effective online Q&A services. Thus, many works have been focused on the prediction of answer quality in online Q&A sites. Jeon, Croft, Lee, and Park (2006) utilizing non-textual features, such as click counts and recommendation history etc., presented a framework to predict the quality of answers. According to their work, answer length has the most effect in quality prediction among all proposed features. Using another set of content-appraisal features, such as informative, readable, polite and so on, Shah and Pomerantz (2010) found that most of those factors did not contribute significantly in explaining the variability of answer qualities. However, when experimenting with automatically extracted features from both the questioners and answerers, the authors then found that contextual information such as the profile of the questioner as well as the answerer may play an important role in determining answer quality. In a more recent study, Harper et al. (2010) noticed that fee based Q&A sites (e.g. Google Answers) tend to receive higher quality answers as compared to those free Q&A sites (e.g. Yahoo!Answers). In addition, sites where anybody can answer questions outperformed sites that depend on experts.

Besides the above-mentioned works targeting quality assessment, there are also some other studies focusing on recommending potential answerers to a given question. A large proportion of these works are related to expert finding, in which the authors constructed expertise models based on topic models (Guo, Xu, Bao, & Yu, 2008; Jurczyk & Agichtein, 2007; Zhu, Cao, Xiong, Chen, & Tian, 2011) or language models (Agichtein, Castillo, Donato, Gionis, & Mishne, 2008; X. Liu et al., 2005; Yandong Liu & Agichtein, 2008) or etc. Although recommending experts to questioners could be beneficial in the context of online Q&A, given the large number
of advise-seeking or conversational questions posted on social Q&A sites (Harper, Moy, & Konstan, 2009), there are some features other than expertise that need to be considered. With this in mind, approaches of question routing have been initiated and developed Balog, Azzopardi, & De Rijke, 2006 in recent Q&A studies. Baichuan Li and King (2010) proposed a question routing framework which considered both the answerer’s expertise and availability when predicting the optimal answerer for a new question. Zhou, Lyu, and King (2012) considered the question routing problem as a classification one, and developed several local and global features. While evaluating their classification results, the authors found that question-user relationship features played a key role in improving the overall performance.

**Social Question and Answering**

Although the terms “community Q&A” and “social Q&A” have been mixed up and used in literatures, according to (Morris, Teevan, & Panovich, 2010a), the experience of asking questions on a Q&A site is different from the experience of Q&A on SNS. First, answers received in SNS can be more credible than those obtained from online Q&A sites. This is because that online Q&A can be viewed as a more impersonal mode of communication, therefore may possess less credibility than information provided by one’s close friends or family members (Goldsmith & Horowitz, 2006). Second, Q&A on SNS includes much fewer potential answers than Q&A on online Q&A sites, since in social Q&A questions posted to one’s social network can only be viewed by one’s direct contacts rather than the whole community. Third, due to the limits of message length on SNS, questions posted on SNS tend to contain fewer characters as compared to questions posted in professional Q&A sites. With all these differences taking into consideration, the Q&A behavior demonstrated on these platforms will also be different.
Questions Asked in Social Q&A

As an emerged concept, the expectation is that social Q&A has potential as an alternative to traditional information seeking tools (e.g. search engines, online catalogs, and professional question-and-answering sites). A number of studies have been conducted from the question asker’s perspective exploring the motivations and patterns of current social Q&A experience.

Morris et al. (2010b) surveyed 624 social network users concerning their reasons for choosing social networks as the platform for Q&A. The results indicated that people search socially primarily due to their trust in friends over trust in strangers. Other than that, specific audience, weak beliefs on search engine performances, and non-urgent information needs also accounted for the reasons people turn to social networks to seek information. To further examine the factors that influence users' adoption of social Q&A, the researchers also conducted another user study (Morris et al., 2010a) and confirmed that seeking information on social networks can provide more personalized answers with higher response quality.

In addition to motivations, other research has been conducted to understand the taxonomy of questions asked on SNS. Through their analysis of 100 question tweets, (Efron & Winget, 2010) found that Twitter users use their social network to satisfy their information needs by asking both factual and impersonal opinion questions to their friends online. Based on their findings, the authors proposed a taxonomy of questions asked under the microblogging environment. Evans and Chi (2008) conducted their study using the taxonomy of traditional search (transactional, navigational, and informational) (Jansen, Booth, & Spink, 2008). The authors presented a social search model of user activities before, during, and after search and proved the value of social interactions in information seeking tasks. Utilizing naturally collected tweets from Twitter, Paul et al. (2011a) assessed whether or not Twitter is a good place for asking
questions. By analyzing question tweets, the authors found that rhetorical questions were the most popular form of questions asked, followed by questions seeking for factual knowledge.

Although attention was given to social Q&A in the past few years, a large proportion of the above mentioned studies were conducted with survey-type data sets, and focused on the intentions behind. Among the few studies characterizing the features of social Q&A through real tweets (Paul, Hong, & Chi, 2011b; Zhao & Mei, 2013), most of the work has been on general questions, including rhetorical ones. Few focused on non-rhetorical ones, since the separation of the two types of questions is algorithmically challenging. Although findings from those studies are informative, given their broad coverage, there still lacks a comprehensive understanding regarding those serious information seeking tweets.

Questions Answered in Social Q&A

Although social search has been studied from the perspective of the question asker, limited work has been done from the standpoint of the answerer. Given the low response rate demonstrated in social search (Sysomos, 2010), many researchers turned their research focus to professional and other Q&A sites to explore the secret of high quality and high quantity response. Adamic, Zhang, Bakshy, and Ackerman (2008) in their work used Yahoo! Answer to understand patterns demonstrated in knowledge sharing activities. By clustering forum categories according to their content characteristics, the authors found a strong association between user’s entropy (the broadness of user’s focus) and the rating of the answers. Harper et al. (2010) conducted field experiments utilizing multiple online Q&A sites and found that fee-driven sites, such as Google Answers received higher quality answers than other online but free sites. They also conclude that answer quality can be affected by both intrinsic factors and extrinsic factors, such as perceived ownership of information and reputation. Besides studies targeting on the answers, there are also
other works studying the quality of responses by identifying the answerer’s level of expertise in certain area (Jurczyk & Agichtein, 2007; X. Liu et al., 2005).

In addition to studies investigating answer qualities, there are also a few works focusing on the perspective of response quantity. Teevan, Ramage, and Morris (2011) conducted a user study in which participants are asked to search the Web while simultaneously posting a question on the same topic to their social network. By comparing the differences between traditional search and social search, the authors found that about half of the subjects received responses from their network before completing their search, which demonstrate the feasibility of using SNS for information seeking. By performing statistical analysis based on real tweets, Paul et al. (2011a) noted that the majority of questions received no response on Twitter. They also found that among those few interrogative tweets with answers, the response rate is strongly related with some of the characteristics of the question askers, such as the size of their networks.

Although studies exploring response qualities on social search are insightful, there are even more basic issues that need to be addressed before I could move to that stage. Despite findings from Teevan et al. (2011) and Paul et al. (2011a) studies, we still know very little about what happens when SNS users see their friend’s or follower’s questions online. What drives them to provide their response for knowledge sharing purpose? What deters them from responding? Given the very low response rate of Q&A on SNS currently, I believe that without a more comprehensive evaluation of factors affecting whether or not a question will be answered, SNS’s power in social information seeking may never be achieved.

Social Q&A Systems

Through online experiments, Nichols and Kang (2012) explored the feasibility of users responding to questions sent by strangers. They found that less than half of the people will answer
to questions posted by strangers; however, they failed to indicate the characteristics of those responders. Pan, Luo, Chi, and Liao (2013) offered a more in-depth analysis on potential answerers by leveraging users’ non-Q&A social activities. Through their analysis of an inter-organizational community Q&A site, they found that some of the non-Q&A features can effectively predict the likelihood of one’s answering of others’ questions. Luo, Wang, Zhou, Pan, and Chen (2014) built a Smart Social QA system based on IBM Connected that recommends both active and inactive users for a given question based on their abilities, willingness, and readiness. The only limitation of their work is that given the differences between organizational community Q&A site and SNS, their framework may meet with some difficulties when generalized to social Q&A context.

To address the gaps in the literature, in the next three chapters, I present three of my research studies, each from one of the perspectives as reviewed above.
Chapter 3

ASK: A Taxonomy of Information Seeking Posts in Social Question and Answering

As understanding users’ information needs is crucial for designing and developing tools to support their social Q&A behaviors, in this chapter, I propose a taxonomy referred to as ASK, which differentiates questions into three types including: accuracy questions, in which people ask for fact or common sense; social questions, in which people ask for coordination or companion; and knowledge questions, in which people seek for personal opinions or advice. Based on the taxonomy, I also develop a predictive model to automatically differentiate these three types of questions, using features from lexical, topical, contextual, syntactic, and interaction and social perspectives. I carry out the experiment of the classification model using questions posted on Twitter. In total, approximately 25,000 question tweets are collected for this study. From those 25,000 questions, I randomly select 3,000 and manually label each of them into accuracy, social, and knowledge. I implement and evaluate multiple classification algorithms with the combination of lexical, topical, contextual, and syntactical features from the question’s aspect, as well as interaction and social features from the answer’s perspective. Through experiments, the proposed classification model proved to be reliable in distinguishing the three types of ASK questions with a classification accuracy of 83.20%.

Background

Taxonomy of social Q&A

In addition to motivations, other research has been conducted to understand the typology of questions asked on SNS. Through an analysis of 100 question tweets, Efron and Winget (2010)
report that Twitter users use their social network to satisfy their information needs by asking both factual and impersonal opinion questions to their friends online. Evans and Chi (Evans & Chi, 2008) conducted their study using the transactional, navigational, and informational taxonomy of traditional search (Jansen et al., 2008). In their study, Morris et al. (2010b) manually labeled a set of questions posted on social networking platforms and identified eight question types in social Q&A, including: recommendation, opinion, factual knowledge, rhetorical, invitation, favor, social connection, and offer. In the set of tweets they analyzed, “recommendation” (29%) and “opinion” (22%) questions accounted for the majority of cases. Differently, Paul et al. (2011b) noticed more rhetorical (42%) questions on Twitter, followed by the categories of factual knowledge (16%) and polls (15%). Ellison, Gray, Vitak, Lampe, and Fiore (2013) labeled a set of 20,000 status updates on Facebook and presented multiple types of mobilization requests beyond information-seeking attempts.

Among the previous works on question typology in a social context, Harper et al. (2010) proposed the most comprehensive classification of questions. By coding questions drawn from three popular community Q&A sites, Harper developed a typology of question types that fall into three “species” of rhetoric, including: deliberative, epideictic, and forensic. Furthermore, the authors proposed some potential alternative taxonomies for future research such as: objective and subjective; past, present, and future; advice, opinion, and factual; and conversational and informational. Gazan (2011) also proposed a need for more appropriate categories of questions that maximize the likelihood of receiving answers.

**Automatic Question Classification**

Most of the above-mentioned studies performed the question classification task manually based on handcrafted rules. There were only a few papers that touch on the problem of automatic
question classification based on machines learning techniques. (Baichuan Li et al., 2011) proposed a cascade approach, which first detected interrogative tweets and then questions revealing real information needs. They relied on both rule-based (as proposed in (Efron & Winget, 2010)) and learning-based approaches for interrogative tweets detection and some Twitter-specific features, such as retweet and mention to extract question tweets. Through their experiment, the authors noticed that the rule-based approach actually outperformed the learning-based method in identifying interrogative tweets. Zhao and Mei (2013) classified question tweets into two categories: tweets conveying information needs and tweets not conveying information needs. They manually labeled 5,000 tweets and built an automatic text classifier based on lexical, POS tagging, and meta-features. With the classifier, they investigated the temporal characteristics of those information-seeking questions. This work is a further step of the above-mentioned studies in the direction of understanding and comprehending the question's intent in social Q&A. It differs in that I develop a practical but also nuanced taxonomy of social Q&A and combine this with automatic classification.

Besides the two works in social Q&A, most of the existing studies on automatic question classification were conducted in the context of community question and answering (community Q&A). Analyzing questions from three popular community Q&A sites, Harper et al. (2009) automatically classified questions into conversational and informational, and reached an accuracy of 89.7% in their experiments. As a result of their analysis, the authors claimed that conversational questions typically have much lower potential archival value than the informational ones. Kucuktunc, Cambazoglu, Weber, and Ferhotosmanoglu (2012) performed the same task based on users’ pre-question behaviors using their toolbar data. Kim, Oh, and Oh (2007) classified questions from Yahoo! Answers into four categories: information, suggestion, opinion, and other. They pointed out that the criteria of selecting the best answer differed across categories. Pal, Margatan, and Konstan (2012) introduced the concept of question temporality
based on when the answers provided on the questions would expire. They labeled questions into four categories, with permanent, long, medium, short, and other temporal durations. Their results showed that question temporality can be detected automatically using question vocabulary and other simple features.

As for the task of question subjectivity identification, Baichuan Li et al. (2011) labeled 987 resolved questions from Yahoo! Answers and explored a supervised learning algorithm utilizing features from both the perspectives of questions and answers to predict the subjectivity of a question. Zhou et al. (2012) proposed an approach to automatically collect training data based on social signals, such as like, vote, answer number, etc., in community Q&A sites. The results of their experiment demonstrated that leveraging social interactions in community Q&A portals could significantly improve the prediction performance. Long Chen, Zhang, and Mark (2012) classified questions from Yahoo! Answers into: subjective, objective, and social. They built a predictive model based on both textual and meta features and co-trained them. Their experimental results showed that co-training worked better than simply pooling these two types of features together. Aikawa, Sakai, and Yamana (2011) employed a supervised approach in detecting Japanese subjective questions in Yahoo! Chiebukuro. Unlike the other studies, they evaluated the classification results using weighed accuracy, which reflected the confidence of annotation.

As such, there are several open questions in this line of research concerning social Q&A, some of which I address in this research.
Research Objectives

With the ultimate goal to better understand the intent of questions posted in social Q&A in order to choose the most appropriate answering strategies for different types of questions, I established the research objectives of this study as follows:

1. *Develop a taxonomy of questions proposed in social Q&A that could be used to assist in selecting the most appropriate answering strategies.*

Regarding the first research objective, I develop the ASK taxonomy to serve as the theoretical groundwork for the types of questions posted in social Q&A based on underlying user intent of the information desired. ASK uncovers three general types of information needs: *Accuracy, Social,* and *Knowledge.* Unlike previous studies (Efron & Winget, 2010; Ellison et al., 2013; Gazan, 2011) which were primarily descriptive, I believe the proposed taxonomy can enhance question answering on SNS, since categorizing questions into ASK types can result in higher response probability and quality via selecting different answering strategies for different types of questions.

The hypothesis of the proposed taxonomic system is that *accuracy* questions tend to be objective and can be answered by machines (i.e., search engines, archived similar question-answer pairs), whereas in contrast the other two types of questions, *social* and *knowledge,* require more diverse replies that rely on personal experiences and opinions, and are preferably answered by human respondents. In addition, compare to the *accuracy* and *knowledge* questions, *social* questions tend to be more “acquaintance-oriented”. So, in order to answer *social* questions, the social Q&A system needs to count more on the questioner’s own friends / followers on SNS; while, for *knowledge* questions, the system can route the inquiries to a larger audience base for more relevant answers.
2. Implement the proposed taxonomy by automatically classifying questions into ASK types and measure the effectiveness of the classification.

To achieve the second research objective, I explore the distinction between the ASK types of questions in the way they are being proposed and answered. To measure the differences, I introduced features from five different aspects, including: contextual, interaction, lexical, syntactic, and topical. These aspects are defined as:

- **Contextual** – the situational elements related to the proposed question
- **Interaction** – the social relationships inherent in the proposed question
- **Lexical** – the words, expressions, and vocabulary used in the posted question
- **Syntactic** – the structure of the vocabulary of the proposed question
- **Topical** – the major subject(s) of the posted question

I then built a prediction model that can reliably distinguish the three types of questions using machine learning algorithms. I evaluated the proposed model using a large sample of questions collected from Twitter.

**Methods**

**Data Collection**

To answer the proposed research questions, using Twitter API, I collected tweets written in English that were posted from September 20 to October 1, 2014 containing at least one question mark and any of the 5WHC questions words, including: who, where, when, what, why, how, and can.. Given the low percentage of information-seeking questions on Twitter (6% as
reported in (Paul et al., 2011a)) and the scope of this study on informational questions only, I further constrained my search query by including a general set of question-signaling hashtags, as introduced in Rzeszotarski and Morris (2014)’s work to filter out as many conversational / non-information-seeking questions (Harper et al., 2009) as possible. Only questions containing at least one of those hashtags was included in the data set. I also removed retweets and questions directed to specific users (tweets with @username) considering the lack of necessity of question routing. This left me with a total of 23,258 information-seeking tweets.

**Taxonomy Creation**

To address my first research objective, I developed a taxonomy that differentiated questions in a way that is beneficial for developing social Q&A tools, such as question routing, answer ranking, and summarization. To be more specific, I created the ASK taxonomy considering both the subjectivity of the question and the scope of potential respondents, and I categorized them into: accuracy, social, and knowledge. By differentiating questions into those three types, ASK can be used to determine whether the collected answers will be ranked based on authority or summarized for quick digestion. Also, ASK can route questions to the appropriate respondents according to their relationship to the questioner.

To test the proposed ASK taxonomy, I randomly sampled 3,000 questions from the collected data set and recruited two human annotators, who are graduate students majoring in Information Sciences and Technology to perform the labeling task based on the developed annotation criteria for *Accuracy*, *Social*, and *Knowledge* questions.

In order to guide the annotation process and to promote continuity between human annotators, the following annotation criteria were adopted:
• **Accuracy Question**: The intent of an accuracy question is to receive answers based on some factual or prescriptive knowledge. The purpose of the question is to receive one or more correct answers instead of responses based on the answerer’s personal experience. This type of question usually looks for facts, definitions, and prescriptive methods on how to do something.

• **Social Question**: The intent of a social question is to request either companionship or coordination support from others. It includes questions searching for someone who shares the same agendas or for someone who can provide physical or emotional assistance.

• **Knowledge Question**: The intent of a knowledge question is to receive responses reflecting the answerer’s personal opinions, advices, preferences, or experiences. It usually has a “survey” purpose, which encourages the audience to provide personal answers.

Additionally, considering also there are questions that do not convey any real information needed, I also asked the annotators to tag those questions, as they do not belong to the ASK types. So, although the posts are presented as a question, these posts are not really information seeking, but instead information providing or rhetorical.

To better illustrate the taxonomy proposed in this study Table 3-1 lists a number of sample questions with accuracy, social, and knowledge intents.
Table 3-1. Annotation criteria for ASK question types and non-information seeking questions.

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Sample Questions</th>
</tr>
</thead>
</table>
| Accuracy               | • When is the debate on UK time? #replytweet  
• Hey gamers. Anyone know how to turn off motion controls for Hyrule Warriors? #HyruleWarriors #wiiu #controllerproblems #help  
• Anyone know how to say "fun smasher" in Spanish? #help |
| Social                 | • Is there somebody who has an ELLO invitation for me? #dtv #Question  
• Who on lemoyn campus has a 5 charger? #replytweet  
• Um somebody want to teach me what you learned in algebra 2 yesterday? #help  
• Who's going to the game tomorrow? #replytweet |
| Knowledge              | • I hate when the bf has not texted me in five days help guys what should I do? #help  
• How should I get my nails done for homecoming?? #replytweet http://t.co/Lj6xpsR1IG  
• I'm in desperate need of a good book series to read. What would you guys recommend? #replytweet  
• Is it worth upgrading to an #iPhone6? #tech #questions |
| Non-Information-Seeking| • I don't really talk to you guys much and it makes me feel bad and antisocial that I don't. Why is that? #ReplyTweet #DMMMe  
• How is it possible to have this much workload during the first month of school? #overloadincredits #help  
• Why does this product suck so much? #ReplyTweet |

With the above annotation criteria, the two human annotators worked on the labeling separately. There were 2,621 of 3,000 questions (87.37%) agreed upon by the two independent coders, indicating the relatively high reliability and generalizability of the ASK taxonomy. Among the 2,621 question tweets, 1,253 (47.81%) were labeled as having a non-information-seeking intent, 475 (18.12%) as accuracy-seeking, 112 (4.27%) as social-seeking, and the rest 781 (29.80%) were labeled as knowledge-seeking.

I also examined the 379 questions with annotation differences and found that the major cause of such disagreement is that without knowing the very specific context of a question, annotators interpreted the questioner’s intent differently. For instance, for the question, “Ok as
we start to evaluate #Obama legacy who is worse him or #Carter? #QuestionsThatNeedAnswers”, one annotator tagged it as knowledge question as it surveys the audience about their opinions regarding Obama, while the other treated this tweet as sarcasm and tagged it as conversational. To accomplish the proposed research objectives, I used only the 475 accuracy, 112 social, and 781 knowledge questions.

**Question Classification Using ASK Taxonomy**

For the second research objective, in this section, I present the design of a model for the automatic prediction of accurate, social, and knowledge questions posted on Twitter.

**Feature Extraction**

First, I introduced the number of features extracted for the purpose of question classification. In total, I identified attributes from five different aspects:

- **Lexical Attributes**

  Given the different information needs, there should be a different usage of lexical terms among questions of distinct types, so, I included lexical features that operate at a word-level, including both n-gram (n = 1, 2, 3) and POS-tagging (part-of-speech tagging) patterns for each question.

  I adopted word-level n-gram features by counting the frequencies of all unigram, bigram, and trigram tokens that appeared in the training data, as they have been proved to be useful in previous work (Aikawa et al., 2011; Zhao & Mei, 2013). Before feature extraction, I lowercased and stemmed all of the tokens using the Porter stemmer (Porter, 1980). I discarded rare terms
with observed frequencies of less than five to reduce the data’s sparsity. This data processing left us with 996 n-gram features.

In addition to the lexical features, I believed that POS-tagging may also help distinguish the three ASK types of questions, as such annotation can provide more context to the words used in the interrogative tweets. To tag the POS of each tweet, I adopted the Stanford tagger (Toutanova, Klein, Manning, & Singer, 2003). Again, I counted the frequencies of all unigram, bigram, and trigram POS patterns that appeared in the training data. In total, I extracted 664 POS-tagging features. This results in a total of 1660 lexical attributes.

• **Topical Attributes**

The topical features related to the subject categorization of each question. The discriminative effect of question topic on question intent has been quantitatively shown by a number of previous studies (Long Chen et al., 2012; Harper et al., 2009; Baoli Li, Liu, Ram, Garcia, & Agichtein, 2008; Z. Liu & Jansen, 2013). To test its power on classifying ASK types of questions, I ask the two human annotators to manually code the 1,368 informational questions into 13 categories, including: Technology, Entertainment, Beauty & Style, Family & Relationship, Home & Garden, Education, Automobile, Food & Drink, Travel, Healthcare, Pets, Social & Culture, and Word & Reference, which are typical categories in many thematic schemes. The inter-rater agreement between the two annotators was 86.11%, which is quite high given 13 possible categories.

• **Contextual Attributes**

For the contextual features, I focused on the two aspects of temporal and spatial attributes of a question. Z. Liu and Jansen (2012) indicated the high percentage of questions containing either explicit temporal or spatial indicators and claimed that contextual features played important
roles in social Q&A. Later studies (Yefeng Liu, Alexandrova, & Nakajima, 2013; Z. Liu & Jansen, 2014; Wang, Chen, Hou, & Chen, 2014) further verified the feasibility and advantages of building real-time social question-answering services on top of microblogging platforms.

**Extracting Spatial Attributes.** To identify the spatial indicators of a question tweet, I used a service called Alchemy API (http://www.alchemyapi.com/), which is a text mining platform based on natural language processing and machine learning algorithms. The Alchemy API allows one to retrieve an entity’s information, such as people, companies, organizations, geographic names, etc., from a paragraph of input text. It has been adopted in a number of previous studies for text mining tasks (Batool et al., 2012; Quercia, Capra, & Crowcroft, 2012). For the purpose of spatial extraction, I only extract entities with the types of “City”, “Continent”, “Country”, “Facility”, and “Region”.

**Extracting Temporal Attributes.** For temporal extraction, I adopted the Stanford temporal tagger (SUTime) (Chang & Manning, 2012). SUTime is a library for recognizing and extracting date, time, and duration values. It is rule-based and extracts explicit temporal expressions based on regular expressions. SUTime exhibits stronger performance compare with other temporal taggers, like HeidelTime (Strötgen & Gertz, 2010) for English text (Sil & Cucerzan, 2014). With SUTime, I identified a list of temporal expressions, such as “today”, “this week”, “2:30am”, “Halloween”, “tomorrow night”, etc., in the data set.

For the contextual attributes, I replaced the extracted exact temporal expressions with `<TIME>` and spatial expressions with `<LOCATION>` in the data set, and rebuilt the classification model with these more general contextual factures.

- **Syntactic Attributes**
The syntactic features measured the writing style of the question at the sentence level and above. The syntactic features that I adopted in this study included: the length of tweets in sentences / clauses, words, and characters, and whether or not the tweet contains a picture. In order to identify tweets containing pictures, I expanded all shortened URLs through a website called LongURL (http://longurl.org/).

- **Interaction Attributes**

  As mentioned earlier, different types of questions may lead to answers with distinct characteristics, so I included the interaction features that recorded the total number of answers received, the unique number of answerers, and the answer length in words for each question.

  In addition, as mentioned above, I hypothesized that questions of different types required different ranges of people to answer. For instance, I believe that even strangers can answer most of the accuracy questions, whereas, social questions usually are implicitly restricted to online friends/followers only; therefore, I retrieved the relationships between the questioners and the answers and included whether or not a stranger answers the question as a feature in the classification model. I also analyzed features including: the length of the answer, the number of overlapped terms in all answers, the arrival time of the first answer, the arrival time of the last answer, and whether or not an answer contains an URL, which are all features indicating the interaction aspects.

**Classification Algorithms and Evaluation Metrics**

With the above features, I next built a multi-class classifier to automatically label questions into the three ASK types: accuracy, social, and knowledge. I trained and tested the
proposed model using a number of classification algorithms implemented in Weka, including: Naïve Bayes, SVM (SMO), and decision tree (J48), using 10-fold cross-validation for each classification experiment.

For evaluation purposes, I used the traditional metrics, including: precision, recall, F1, and accuracy as they have also been adopted in many other studies (Bian et al., 2008; Castillo, Mendoza, & Poblete, 2011; Zhou et al., 2012). I also adopted the majority induction algorithm, which predicts the majority class in the data set as a baseline model to interpret the classification results and evaluate the classification method. With this approach, the data set got a baseline accuracy of 0.571 as 781 tweets were tagged as knowledge-seeking among the overall 1,368 informational seeking questions.

Results

In this section, I address my second research question by understanding the classification results of question types.

Classification Results using Lexical Features

Due to the large number of lexical features extracted, I evaluated the classification accuracies along the number of features selected. Figure 3-1 shows the feature selection results using the algorithm of information gain (Y. Yang & Pedersen, 1997) implemented in Weka (Hall et al., 2009). Information gain measures the variation in entropy when a feature is present versus when it is absence and is frequently employed as a feature selection method in the field of machine learning.
As can be seen from Figure 3-1, either too few or too many features would result in a decrease of prediction accuracy. In addition, SMO outperformed the other two methods in the question classification process using lexical features with 0.820 accuracy, 0.823 precision, 0.820 recall, and 0.821 F1-measurement. I also noted that the accuracy of 0.820 was much better than the majority class baseline of 0.571, which validated the possibility of automatically detecting question types using lexical features only. Table 3-2 presents the classification results of all three methods.

Table 3-2. Performance of the lexical-based classifier.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMO</td>
<td>0.823</td>
<td>0.820</td>
<td>0.820</td>
<td>0.821</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.748</td>
<td>0.747</td>
<td>0.747</td>
<td>0.747</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.768</td>
<td>0.763</td>
<td>0.763</td>
<td>0.765</td>
</tr>
</tbody>
</table>

I adopted the method of information gain to identify the most informative and relevant features of each question type. Table 3-3 shows the top 10 discriminative word features. From
Table 3-3, I noticed that about half of the accuracy and knowledge tweets contained the question word “what”; whereas, 90% of the social questions started with the word “who”, indicating their companion-seeking nature. Example accuracy and knowledge questions containing “what” and social questions containing “who” include: “What does #tsibip mean? #twoogle”, “what is a
good music downloader app? #replytweet” and “who is going to the eagles tomorrow and wants
tailgate? #ReplyTweet”. I also discovered that, while compared with the other two types of
questions, knowledge-seeking tweets asked more about locations. These questions tended to
include more contextual information, using the word “I” in 48.34% of 782 cases. Typical
knowledge questions are: “If I ever were to replace my beloved Pentax K200D (and I will have
to), what should I get next? Another Pentax? Canon? Nikon? #question”.

Table 3-3. Top 10 n-gram features of each ASK question type along with the informational gain.

<table>
<thead>
<tr>
<th>Word</th>
<th>Accuracy (%)</th>
<th>Social (%)</th>
<th>Knowledge (%)</th>
<th>Information Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>who</td>
<td>7.98</td>
<td>90.90</td>
<td>6.26</td>
<td>0.21</td>
</tr>
<tr>
<td>I</td>
<td>12.18</td>
<td>10.00</td>
<td>48.34</td>
<td>0.12</td>
</tr>
<tr>
<td>what</td>
<td>45.59</td>
<td>2.73</td>
<td>49.49</td>
<td>0.06</td>
</tr>
<tr>
<td>should</td>
<td>0.63</td>
<td>0.00</td>
<td>13.68</td>
<td>0.06</td>
</tr>
<tr>
<td>good</td>
<td>0.00</td>
<td>6.36</td>
<td>11.64</td>
<td>0.05</td>
</tr>
<tr>
<td>who is going</td>
<td>0.21</td>
<td>15.45</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>where</td>
<td>7.77</td>
<td>1.82</td>
<td>22.25</td>
<td>0.04</td>
</tr>
<tr>
<td>what time</td>
<td>6.72</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>me</td>
<td>4.20</td>
<td>29.09</td>
<td>5.63</td>
<td>0.03</td>
</tr>
<tr>
<td>want</td>
<td>0.63</td>
<td>18.19</td>
<td>3.20</td>
<td>0.03</td>
</tr>
</tbody>
</table>

**Classification Results using Topical Features**

In addition to looking at the lexical features, I also examined the topical categories
annotated by the two human coders. As mentioned earlier, I divided the 1,368 informational
seeking questions into 15 categories according to their topics. Table 3-4 shows the 15 topics and
the percentage of questions belonging to each category. The topical categorization results were consistent with previous studies (Z. Liu & Jansen, 2013; Morris et al., 2010b), suggesting a predominance of “Technology”, “Entertainment”, and “Education” questions; however, I also noticed a large percentage of “Beauty and Style” related questions in the collected data set with the questioners asking their friends for opinions on their appearance that were not discussed in previous works.

While analyzing the percentage of accuracy, social, and knowledge questions within each topical category, I identified that the majority of “Beauty & Style” (89.53%), “Food” (84.88%), “Health” (92.86%), “Family & Relationship” (100%), and “Home & Garden” (80.00%) questions sought other’s opinions and advice, whereas 96.97% of “Word & Reference”, 63.21% of “Sports”, 67.27% of “Society & Culture”, and 53.38% of “Education” questions were of an objective nature.

Table 3-4. Fraction of ASK questions within each topical category.

<table>
<thead>
<tr>
<th>Question Topic</th>
<th>Total (%)</th>
<th>Example Question</th>
<th>Accuracy (%)</th>
<th>Social (%)</th>
<th>Knowledge (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>22.15</td>
<td>Can anyone help me figure out why I’m not able to load any links from a tweet? A &quot;loading error&quot; keeps popping up #help</td>
<td>43.23</td>
<td>3.30</td>
<td>53.47</td>
</tr>
<tr>
<td>Entertainment</td>
<td>16.45</td>
<td>Does anyone know when Vampire Diaries starts?!?! #replytweet</td>
<td>33.33</td>
<td>10.23</td>
<td>56.44</td>
</tr>
<tr>
<td>Beauty &amp; Style</td>
<td>13.96</td>
<td>Ok ladies I need help!! Navy dress...what color shoes? I have red, fuchsia, silver, gold...no nude though...#replytweet</td>
<td>5.76</td>
<td>4.71</td>
<td>89.53</td>
</tr>
<tr>
<td>Education</td>
<td>9.72</td>
<td>Anyone in my law class know what part three of the assignment is??? #replytweet</td>
<td>53.38</td>
<td>18.05</td>
<td>28.57</td>
</tr>
<tr>
<td>Sports</td>
<td>7.75</td>
<td>What time is the soccer game tonight?? #replytweet</td>
<td>63.21</td>
<td>19.81</td>
<td>16.98</td>
</tr>
<tr>
<td>-----------------</td>
<td>------</td>
<td>--------------------------------------------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Food</td>
<td>6.29</td>
<td>Am hungry but i do not know where to go!! Any suggestions?? #help</td>
<td>11.63</td>
<td>3.49</td>
<td>84.88</td>
</tr>
<tr>
<td>Society &amp; Culture</td>
<td>4.02</td>
<td>What is y'all's opinion on breast feeding in public? #AskTwitter</td>
<td>67.27</td>
<td>1.82</td>
<td>30.91</td>
</tr>
<tr>
<td>Travel</td>
<td>3.66</td>
<td>Really want to visit the MiddleEast next year, as a Brit, where would people recommend? #UAE #SaudiArabia #Dubai #Qatar #Kuwait? #help</td>
<td>24.00</td>
<td>20.00</td>
<td>56.00</td>
</tr>
<tr>
<td>Family &amp; Relationship</td>
<td>3.58</td>
<td>How do guys usually act around a girl they're interested in? Trying to decode someone. #ReplyTweet</td>
<td>0.00</td>
<td>0.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Business</td>
<td>3.29</td>
<td>How do I get a job at Ukhozi FM? #Twoogle</td>
<td>37.78</td>
<td>17.78</td>
<td>44.44</td>
</tr>
<tr>
<td>Health</td>
<td>3.07</td>
<td>What should I take for a sore throat &amp; a runny nose? #replytweet</td>
<td>4.76</td>
<td>2.38</td>
<td>92.86</td>
</tr>
<tr>
<td>Word &amp; Reference</td>
<td>2.41</td>
<td>What is #LRT? #twoogle</td>
<td>96.97</td>
<td>0.00</td>
<td>3.03</td>
</tr>
<tr>
<td>Home &amp; Garden</td>
<td>2.19</td>
<td>How do you clean washing machines? Mine smells a bit musky! #housework #help</td>
<td>13.33</td>
<td>6.67</td>
<td>80.00</td>
</tr>
<tr>
<td>Pets</td>
<td>0.88</td>
<td>I am adopting myself a dog... where is the best place to get a rescue? #asktwitter</td>
<td>41.67</td>
<td>0.00</td>
<td>58.33</td>
</tr>
<tr>
<td>Automobile</td>
<td>0.58</td>
<td>Does any1 in Tampa kno where to get their car detailed ?? #replytweet</td>
<td>12.50</td>
<td>12.50</td>
<td>75.00</td>
</tr>
</tbody>
</table>

Table 3-5 illustrates the classification results using topical categories. I noticed that although the topic-based classifier outperformed the baseline, its performance was much weaker than the lexical-based model.
Table 3-5. Performance of the topical-based classifier.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMO</td>
<td>0.607</td>
<td>0.667</td>
<td>0.667</td>
<td>0.622</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.604</td>
<td>0.664</td>
<td>0.664</td>
<td>0.620</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.607</td>
<td>0.667</td>
<td>0.667</td>
<td>0.622</td>
</tr>
</tbody>
</table>

Classification Results using Contextual Features

In total, with the Alchemy API and SUTime, I identified 107 (7.82%) questions in the data set with explicit spatial mentions, and 209 (15.28%) questions with temporal constraints, for a total of 316 context attributes. In Table 3-6, I depicted the distribution of these questions according to their intent. I found that compared with accuracy and knowledge questions, the social category contained significantly more location- (13.6%) and temporal-specific (34.5%) inquiries. Example of such questions include: “WHO wants to come with me to see Bassnectar at Madison Square Garden on oct. 4?????? #replytweet”, and, “Who wants to go to Peru with me next summer and climb Machu Picchu? #seriousquestion http://t.co/l80Dd0emo3”.

Table 3-6. Distribution of ASK questions with and without location or temporal mentions.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Social</th>
<th>Knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>With Location</strong></td>
<td>27 (5.7%)</td>
<td>15 (13.6%)</td>
<td>65 (13.6%)</td>
</tr>
<tr>
<td><strong>Without Location</strong></td>
<td>449 (94.3%)</td>
<td>95 (86.4%)</td>
<td>717 (91.7%)</td>
</tr>
<tr>
<td><strong>With Time</strong></td>
<td>64 (13.4%)</td>
<td>38 (34.5%)</td>
<td>107 (13.7%)</td>
</tr>
<tr>
<td><strong>Without Time</strong></td>
<td>412 (86.6%)</td>
<td>72 (65.5%)</td>
<td>675 (86.3%)</td>
</tr>
</tbody>
</table>

In addition, I also conducted cross-tabular analysis with Chi-Square test to estimate the dependency between location or temporal mention and question topic. Table 3-7 and Table 3-8
report the top five question topics results, including location and temporal mentions. I observed that among all topical categories, the travel-related questions contained the most temporal and spatial constraints, followed by the topics “food”, “business”, and “entertainment”. Besides, more than 30% of sports-related questions were characterized with temporal mentions, such as, “who do we play Friday? #replytweet” and, “Anyone know who is going to start for the October 10th game? #ColombiavsSalvador #replytweet”. Lastly, 16.4% of “society & culture” questions were targeted at certain location or area. Examples include: “Alright, how to get a marriage certificate attested in Dubai? #Help”, and, “When immigrating to Canada do they transfer your drivers license over or is there a mandatory test first? #askingforafriend #makeitharder”.

Table 3-7. Top 5 question topics by percentage of location mentions.

<table>
<thead>
<tr>
<th>Topic</th>
<th>With Location</th>
<th>Without Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Travel</td>
<td>28 (56.0%)</td>
<td>22 (44.0%)</td>
</tr>
<tr>
<td>Society &amp; Culture</td>
<td>9 (16.4%)</td>
<td>46 (83.6%)</td>
</tr>
<tr>
<td>Food</td>
<td>12 (14.0%)</td>
<td>74 (86.0%)</td>
</tr>
<tr>
<td>Business</td>
<td>4 (8.9%)</td>
<td>41 (91.1%)</td>
</tr>
<tr>
<td>Entertainment</td>
<td>17 (7.6%)</td>
<td>208 (92.4%)</td>
</tr>
</tbody>
</table>

Table 3-8. Top 5 question topics by percentage of temporal mentions.

<table>
<thead>
<tr>
<th>Topic</th>
<th>With Time</th>
<th>Without Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Travel</td>
<td>16 (32.0%)</td>
<td>34 (68.0%)</td>
</tr>
<tr>
<td>Sports</td>
<td>32 (30.2%)</td>
<td>74 (69.8%)</td>
</tr>
<tr>
<td>Business</td>
<td>10 (22.2%)</td>
<td>35 (77.8%)</td>
</tr>
<tr>
<td>Food</td>
<td>18 (20.9%)</td>
<td>68 (79.1%)</td>
</tr>
<tr>
<td>Entertainment</td>
<td>46 (20.4%)</td>
<td>179 (79.6%)</td>
</tr>
</tbody>
</table>

By replacing the spatial and temporal expressions with the corresponding annotations (<TIME> and <LOCATION>), I rebuilt the classification model using the word level n-gram (n
= 1, 2, 3) features. Again, I removed words and phrases with frequencies less than five to reduce feature sparsity. Table 3-9 compares the classification performance of the n-gram features with and without replacement of location and temporal mentions. The general contextual features improved the performance over just the n-gram features, although not significantly.

Table 3-9. Performance of the general contextual-based classifier.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>With Temporal and Spatial Replacement</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMO</td>
<td>0.805</td>
<td>0.803</td>
<td>0.803</td>
<td>0.804</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.765</td>
<td>0.765</td>
<td>0.765</td>
<td>0.765</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.768</td>
<td>0.765</td>
<td>0.765</td>
<td>0.766</td>
</tr>
<tr>
<td>Without Temporal and Spatial Replacement</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMO</td>
<td>0.805</td>
<td>0.801</td>
<td>0.801</td>
<td>0.802</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.764</td>
<td>0.764</td>
<td>0.764</td>
<td>0.764</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.769</td>
<td>0.765</td>
<td>0.765</td>
<td>0.766</td>
</tr>
</tbody>
</table>

Classification Results using Syntactic Features

Table 3-10 illustrates the classification results using syntactic features, including: number of clauses, words, characters, and whether or not the tweet contains a picture, as a form of information seeking and sharing (Jansen, 2008). Again, the syntactic-based classifier outperformed the majority-voted baseline, although its predictive power is limited.

Table 3-10. Performance of the syntactic-based classifier.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMO</td>
<td>0.327</td>
<td>0.572</td>
<td>0.572</td>
<td>0.416</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.558</td>
<td>0.621</td>
<td>0.621</td>
<td>0.577</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.563</td>
<td>0.577</td>
<td>0.577</td>
<td>0.561</td>
</tr>
</tbody>
</table>

Upon further review of question length (as shown in Table 3-11), it is observed that on average, knowledge questions were significantly longer than the accuracy and social ones on all
three levels. Through my subsequent investigation on the content of questions, I noted that knowledge questions tended to use more words to provide additional contextual information about the questioner’s information needs. Examples of such questions include: “Any ideas where I can get some keepsake trunks from? Want something special to store memorable bits for each member of the family. #help”, “What kind of laptop should I get for college work and possibly some online gaming with B? #replytweet #help #laptop #gaming”.

Table 3-11. Question length across ASK question types.

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Num Clauses</th>
<th>Num Words</th>
<th>Num Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std Dev</td>
<td>Mean</td>
</tr>
<tr>
<td>Accuracy</td>
<td>1.321</td>
<td>0.696</td>
<td>11.286</td>
</tr>
<tr>
<td>Social</td>
<td>1.318</td>
<td>0.753</td>
<td>11.009</td>
</tr>
<tr>
<td>Knowledge</td>
<td>1.669</td>
<td>0.991</td>
<td>14.670</td>
</tr>
</tbody>
</table>

In total, I extracted 67 (4.90%) questions from the data set containing pictures. Through my analysis, I found that 4.8% of accuracy, 1.8% of social, and 5.4% of knowledge questions included pictures that were related to the post's content; however, based on the analysis, no statistical difference was noted across the three question categories ($\chi^2 = 2.637$, df=2, p=0.267 > 0.05).

Ensemble Classification Results using Features from the Question’s Perspective

So far, I performed the classification tasks along each individual dimension from the question’s perspective, one at a time. In this section, I constructed an ensemble classifier by using all of those features simultaneously. I reported the ensemble classification results in Table 3-12, from which one can see an improvement over all of the aforementioned models using
features from just a single dimension. The ensemble classifier correctly classified 72.1% of the accuracy questions, 74.5% of the social questions, and 79.2% of the knowledge ones, achieving an average accuracy of 83.2%.

Table 3-12. Performance of the ensemble-based classifier.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMO</td>
<td>0.836</td>
<td>0.832</td>
<td>0.832</td>
<td>0.833</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.772</td>
<td>0.770</td>
<td>0.770</td>
<td>0.771</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.767</td>
<td>0.763</td>
<td>0.763</td>
<td>0.765</td>
</tr>
</tbody>
</table>

Classification Results using Features from the Answer’s Perspective

In addition to the features from the question’s perspective, I also studied the interaction and social patterns from the answer’s aspect. In total, among all 1,368 questions, 574 questions received at least an answer. The descriptive statistics for answered questions were shown in Table 3-13. Consistent with previous studies (Z. Liu & Jansen, 2013; Paul et al., 2011a), less than half of questions received replies. On average, knowledge questions got the highest number of answers from the most number of unique respondents.

Table 3-13. Descriptive Statistics of answered questions across ASK types.

<table>
<thead>
<tr>
<th>Question Type</th>
<th>#Answered (%)</th>
<th>Avg #Answers</th>
<th>Avg #Unique Answerers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std Dev</td>
<td>Mean</td>
</tr>
<tr>
<td>Accuracy</td>
<td>182 (38.32%)</td>
<td>4.39</td>
<td>3.94</td>
</tr>
<tr>
<td>Social</td>
<td>44 (39.29%)</td>
<td>3.95</td>
<td>2.65</td>
</tr>
<tr>
<td>Knowledge</td>
<td>348 (44.56%)</td>
<td>5.16</td>
<td>4.38</td>
</tr>
</tbody>
</table>
As for the analysis on the followship between the questioner and the respondent, among all 1,171 unique questioner-answerer pairs in the data set, 886 (75.66%) of the follow relations were reciprocal; 108 (9.22%) were one-way, and 143 (12.21%) were not following each other. The number of reciprocal following relations in my collection is relatively high, which is consistent with the 70%-80% and the 36% rates as reported in (Paul et al., 2011a; P. Zhang, 2012). I only treated individuals with reciprocal followship as “friends” and the rest as “strangers”. Chi-square test examining the dependency between the questioner-respondent friendship and the answered question type revealed a significant trend between the two variables ($\chi^2 = 7.74$, $p = 0.02 < 0.05$). I found that “strangers”, interestingly, were more likely to answer knowledge questions (29.6%) than “friends”, perhaps support of the concept of the theory of weak ties (Granovetter, 1983). However, this was unexpected given previous work (Morris et al., 2010b) showed that people surveyed claimed that they preferred to ask subjective questions to their friends for tailored responds. However, the observed behavior in the data set does not support this claim. One reason for this could also be that compared to accuracy (20.3%) and social questions (15.9%), knowledge questions, such as: “Need new phone what should I get? #help”, require less expertise and time investment, so that could be a better option for strangers to offer their help.

In addition to examining the relationship between the type of followship and the answered question type, Kruskal-Wallis tests were also performed on the average answer length. The results were significant with answers to the knowledge questions containing the most words ($\text{Mean}_w = 8.73$, $\text{StdDev}_w = 5.48$) and characters ($\text{Mean}_c = 48.58$, $\text{StdDev}_c = 29.24$), and answers to the social questions containing the least ($\text{Mean}_w = 6.82$, $\text{StdDev}_w = 4.53$; $\text{Mean}_c = 34.96$, $\text{StdDev}_c = 24.69$) ($p = 0.00 < 0.05$). I plotted the cumulative distribution of answer length across question types in Figure 3-2.
Figure 3-2. Distribution of question length on word and character levels across ASK types.

Considering the real time nature of social Q&A, I also looked at how quickly the three different types of questions received responses. I adopted two metrics in this study to measure the response speed: (1) the time elapsed until receiving the first answer, and (2) the time elapsed until receiving the last answer. In Figure 3-3, I plotted the empirical cumulative distribution of response time in seconds using both measurements. I logarithmically transformed the response time given its logarithmic distribution.
In general, about 70% of questions in the data set posted on Twitter received their first answer within an hour, no matter their question types (73.10% accuracy, 63.64% social, and 67.53% knowledge questions). From Figure 3-3, it took slightly longer for individuals to answer knowledge questions than the accuracy and social ones. The Kruskal-Wallis result also revealed a significant difference in the arrival time of the first answer across question types (p = 0.00 < 0.05). In addition to the first reply, I also adopted the arrival time of the last answer to imply the temporality of each question. As defined in Pal, Margatan, et al. (2012)’s work, question temporality is “a measure of how long the answers provided on a question are expected to be valuable”. Again, the Kruskal-Wallis result demonstrated significant between-group differences on the arrival time of the last answer (t = 3.76, p < 0.05), with knowledge questions having the longest temporality and the accuracy ones having the least. An example of accuracy questions with short temporal durations is: “When is the game today? #replytweet”. The Kruskal-Wallis
result was also significant ($p = 0.00 < 0.05$) in that almost no *social* question contained any URL that pointed to external sources.

Table 3-14 illustrates the classification results using the interaction features from the answer’s perspective.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMO</td>
<td>0.836</td>
<td>0.832</td>
<td>0.832</td>
<td>0.833</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.772</td>
<td>0.770</td>
<td>0.770</td>
<td>0.771</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.767</td>
<td>0.763</td>
<td>0.763</td>
<td>0.765</td>
</tr>
</tbody>
</table>

### Discussion and Design Implications

From literature on identifying information-seeking questions in social Q&A (Efron & Winget, 2010; Ellison et al., 2013; Paul et al., 2011b; Zhao & Mei, 2013), I conducted this research to investigate the intent of questions asked in social Q&A. Unlike taxonomies proposed in previous studies, ASK was developed to cope with information seeking intents that can be answered with different strategies. I claimed that *social* questions are targeted mainly at one’s online friends / followers rather than strangers; *knowledge* questions are for survey purposes and require as many responses as possible; whereas, computer generated responses could satisfy accuracy-seeking questioners. The proposed ASK taxonomy proved to be reliable with a relatively high inter-annotator agreement.

In addition, this study also proved the feasibility of automatically classifying questions by their underlying intent. I assessed the effectiveness of the classifier and demonstrated its reliability in distinguishing the ASK types of questions with a classification accuracy of 83.20%. Through more in-depth analyses of all three types of questions, I found that, first, regarding the
phrasing of questions, my result implied that contextual restrictions (e.g., time and location) were imposed more often on knowledge and social questions. In addition, my results revealed that knowledge questions experienced longer time-lags in getting their initial answers and also tended to expire in shorter durations. Lastly, I demonstrated that even though individuals prefer to ask subjective questions to their friends for tailored responds, it turned out that in reality, knowledge questions were being responded to more by strangers than accuracy and social questions. I thought this gap between the ideal and reality imposed a design challenge in maximizing the personalization benefits from strangers in social Q&A.

My study on the ASK types of questions has importance for both researchers and practitioners for several reasons. From the theoretical point of view, the proposed taxonomy, ASK, can be adopted as a theoretical groundwork for future studies on social Q&A. Also, the predictive model enables automatic identification of question types and can be used to facilitate future studies in social Q&A on large scales. In terms of design implications, the ASK taxonomy allows the practitioners to understand the distinct intentions behind all three types of questions, and then to design and develop better social Q&A systems to support question answering on SNS. To be more specific, the analysis results in this study reveals the inadequacy of using one single answering strategy (i.e., searching through the archived question-answer pairs, question routing) to serve all social Q&A questions, and suggests a hybrid approach in future social Q&A systems. In Figure 3-4, I illustrate a possible design of such hybrid Q&A system relying on the ASK taxonomy. Under that condition, once questions are asked, the ASK classifier will automatically categorize the questions into three types. Considering the factual nature and short duration of those accuracy questions, the system will try to generate replies based on archived similar question-answer pairs, without human intervention; whereas, given the survey nature of knowledge questions and stranger’s interests in answering them, one could develop an algorithm to route those subjective questions to appropriate respondents based on their locations and past
experiences. Lastly, considering the “acquaintance-oriented” nature of those social questions, the system could leave them for one’s online friends / followers to answer.

Figure 3-4. Hybrid social Q&A system leveraging on ASK taxonomy for question routing.

Conclusion

In this research, I present a taxonomy called ASK that differentiated online information seeking questions into three types: accuracy, social, and knowledge, according to their different scope of targets and different requirements for answers. Based on the ASK taxonomy, I developed and implemented a predictive model based on features constructed from lexical, topical, contextual, syntactic, and interaction perspectives using machine learning techniques. The automated method proved to be effective in classifying ASK types of questions proposed in social Q&A, with a high classification accuracy of 0.832. I further proposed design implications
replying on the ASK taxonomy and associated automated classification method, proposing a hybrid social Q&A system with different routing schemes based on ASK question types. I believe that my study offers valuable insights into the future development of social Q&A systems or tools that can better understand the information seeking needs of the questions posted in social Q&A, as well as to make good use of the automated classification method to adopt different strategies in answering different types of questions.
Chapter 4

Predicting the Response Probability in Social Question and Answering on Sina Weibo

Although social Q&A has a number of advantages over traditional information seeking methods, since it is originally designed as a platform for information dissemination rather than a tool for Q&A, it also has the defect of no guaranteed responses. According to previous studies (Paul et al., 2011a) only 18% of questions posted on Twitter ever receive a response, which is only 1/3 of the response rate of professional Q&A sites, such as Yahoo!Answers. This might be due to various reasons, including the unique design of SNS, as well as the huge volume of data generated on it.

To solve the problem of low response rate, in this work I examine the response rate in social Q&A. I propose in total 24 factors from both the question and the questioner’s aspects to perform the analysis. I limit the scope to extrinsic factors only, given the existence of literature focusing on factors from both social and cognitive perspectives in knowledge sharing (Chow & Chan, 2008; Lin, Hung, & Chen, 2009). For the analysis, I retrieve over 20 thousand real-world questions from Weibo.

The contributions of this study are mainly in two folds: First, although a number of studies have been published regarding people’s behavior in social Q&A, few of them have addressed the problem of low response rate. Besides, to the best of my knowledge, this work is one of the first published studies of social Q&A on the Chinese site, Weibo, although I also expect the findings of this work to inform behaviors on other social network mediums too. Second, I analyze a set of extrinsic factors that are likely to influence the question response rate in social Q&A. By identifying those factors, I could then incorporate them into the design of social Q&A tools or services. For instance, by predicting the number of responses of questions yet to be
answered, I could then route questions with low or no predicted answer rate to more people on SNS for help.

**Background**

Defined as the process of finding information online with the assistance of social resources (Morris et al., 2010b). By making use of all possible social interactions online, social Q&A surpasses the traditional information seeking techniques (e.g. search engine and online databases etc.) with more personalized search experience (Evans & Chi, 2008). Jansen, Zhang, Sobel, and Chowdury (2009) in their work examining Twitter as a mechanism for word-of-mouth advertising reported that 11.1% of the brand-related tweets were information-providing, while 18.1% were information-seeking. Li et al. (Baichuan Li et al., 2011) revealed that there were about 11% of general tweets containing questions (similar to 13% reported in Efron and Winget’s work) and 6% of tweets having information needs.

However, since SNS is not originally designed as a platform for Q&A, not all questions posted on it are guaranteed to receive responses. Paul et al. (2011a) noted that the majority of questions posted on Twitter received no response. They also observed that distinct question types lead to different response rates. For instance, they found that some rhetorical questions received a relatively large number of replies as compared to personal and health-related questions. In addition, the response rate is strongly related with some of the characteristics of the question askers, such as the size of their networks. Nichols and Kang (2012) further confirmed this finding in their online experiment of sending questions to strangers for help. In their results, only less than half of the questions received responses from strangers. Z. Liu and Jansen (2013) studied the social Q&A responses posted on Sina Weibo, the largest Chinese microblogging site. They found that the topic of a question could effectively affect its response rate. For instance, they noticed
that questions of the topics of “Entertainment”, “Society”, “Computer”, etc. received fewer responses as compared to questions from the other categories. Lampe, Gray, Fiore, and Ellison (2014) studied a set of public status updates posted to Facebook and found that mobilization requests got more responses than other kinds of posts. Additionally, they noted that the response speed was affected by the type of support requested.

**Research Objectives**

Given the relatively low response rate in social Q&A, we need a more comprehensive evaluation of what drives individual intentions on helping others in information seeking. Specifically, I aim to focus on two research objectives:

1. *Identify the extrinsic factors that are likely to influence the question response rate in social Q&A.*

2. *Build a model to predict the question response probability using the extrinsic features.*

More formally, I describe the research objective as: given a set of questions posted on SNS $q_1, q_2, q_3, \cdots, q_n$, and the number of answers they received $a_1, a_2, a_3, \cdots, a_n$, predict if a new question $q_{n+1}$ will be answered by someone or not with answers $a_{n+1}, a_{n+2}, a_{n+3}, \cdots, a_{n+r}$. 
Methods

Data Collection

To study the response probability in social Q&A, I collected data from China’s largest microblogging site, Sina Weibo (http://weibo.com/). Sina Weibo attracted over 600 million registered accounts by September 2014 (Bai, 2014), which accounted for 93.60% of the total Internet users in China. Each month, more than two billion statuses are posted on Weibo.

Although Sina Weibo essentially adopted similar operating concepts and functions as Twitter, it differs from Twitter in some aspects. First, since it was launched in 2009, Sina Weibo adopted use of threaded comments, which Twtiter used until its re-design in 2013. As can be seen in Figure 4-1, similar to the commenting system in traditional blogs, comments to the same post on Weibo are shown in a chronological thread. The threaded comments design adds a new dimension of interactivity and engagement to microblogging, thus makes it a better source for studies focusing on the behaviors of social response. Second, due to the fact that Chinese characters are logograms rather than phonograms, the same number of Chinese characters can convey more information than English letters; therefore, with the same 140-character limit, Weibo users can post much more elaborate questions and answers compared to Twitter users.

Considering the huge percentage of conversational questions posted on SNS (Baichuan Li et al., 2011), I adopted a keyword-based method for data collection. I searched for all questions posted from May 1 to September 30, 2014 that contained at least one of the information seeking keywords, including “anybody can tell” (谁能告诉), “please recommend” (求推荐), and “almighty Weibo” (万能的微博, a common Weibo terminology used to express one’s desperate needs for help and information). The modified snowball sampling method was used to identify the keywords used for extracting questions. To control the potential effect of the response time on
the number of answers received, I intentionally collected the questions and answers a week after they were posted. Given the seven day gap and that 97% of tweet replies to questions happen within an hour (J. Yang & Counts, 2010; J. Yang & Wei, 2009), I believed the varied response period would not impact the number of replies collected.

By using the pre-defined search queries, I collected 126,071 questions from 106,367 unique Weibo accounts in five months. While observing the collected data, I noticed that there were many questions with identical content. After a detailed review of those questions, I assumed this was an indication of either spam accounts, which generate posts automatically by copying others or by advertisement campaigns, which contain non-information-seeking questions. To preserve the quality of the study’s data, I decided to delete these questions. To ensure the proposed model’s generalizability, I also removed questions from individuals with more than...
10,000 followers or less than 20 followers. These deletions left us with 62,106 unique questions from 57,699 users. For each question in the data set, I collected its content, posting time, number of comments, and questioner profile via the Sina Weibo API. I also obtained up to 135 latest messages from each questioner in order to measure their posting styles as well as their social attractiveness and engagement.

**Feature Engineering**

To predict the response rate of questions posted on Sina Weibo, I introduced three different types of features: the question features, the user features, the social activity features. The question features captured the linguistic and syntactic properties of each question. The user features described the questioner characteristics. The social activity features depicted the social attractiveness of the questioner.

Next, I introduced each of the three types of features in detail.

**Question Features**

The content and syntactic features consider the characteristics of a question, either Weibo-dependent or Weibo-independent, including:

1. **Question Length:** Previous work (Choi, Kitzie, & Shah, 2012) suggested that question length can be used to measure question complexity. A longer question requires more cognitive effort to process the information needed, which might in turn lead to response failure; however, I believe that the length of a question also reflects how detailed a question is explained. In that sense, the longer the question is, the better the respondents can understand the questioner’s
need and can thus provide their help. So, I assume that the question length in both characters and clauses are positively correlated with the response rate.

2. **Urgency:** Request urgency has proven to be positively linked to increased response rate in past studies. To measure the urgency of a question, I included three measurements: (a) urgency words, (b) repeated punctuations, and (c) repeated interjections. Urgency words signify the requirement of immediate action or attention (Hellier, Edworthy, Weedon, Walters, & Adams, 2002). Individuals are more likely to respond to urgent warnings created by using urgent words and high signal intensity, as indicated in Baldwin (2011)’s study. Likewise, repetition regardless of length limitations, indicates its importance in communicating one’s social meanings to the others (Kalman & Gergle, 2010; Suh, Hong, Pirolli, & Chi, 2010). Repeated punctuations and interjections on SNS are a way users emphasis their anxieties and emotions and to attract others’ attentions in the absence of verbal communication.

To extract the urgency features, I analyzed all questions collected by comparing them with predefined lists of urgency words (e.g. “urgent”, “wait on line”, etc.), repeated punctuations (e.g. !!!, ??, !!, etc.), and 23 Chinese interjections (e.g. “啊”, “呀”, “哇”, “吧”, etc., all have no actual meaning). All predefined patterns identified above were based on the observed frequencies of their occurrence based on my analysis of user behavior on Weibo and on use of urgency observed in prior work (Hellier et al., 2002). I used Ansj (https://github.com/NLPchina/ansj_seg), a Chinese word segmentation tool, to segment words and to remove punctuation. I expected a positive correlation between the usage of urgency expressions and the question response rate.

3. **Gratitude:** As a motivation for pro-social behaviors, prior literatures (Bartlett & DeSteno, 2006; Tsang, 2006) have affirmed the significant effect of gratitude expressions on social exchange. It is believed that through both agentic and communal mechanisms, gratitude expressions can enhance the helper’s feelings of self-efficacy and social worth (Bartlett &
DeSteno, 2006), and thus encourage them to engage in pro-social behaviors (Grant & Gino, 2010). I hypothesize that questions with gratitude expressions have higher probabilities of receiving responses. Again, a predefined list containing 14 gratitude (e.g. “thanks”, “feel grateful”, etc.) and reciprocation words (“pay it back”, “return the favor”, etc.) was adopted to determine the existence of gratitude expressions in all questions collected.

4. **Posting Time Period**: Previous studies addressed the significant variation in Internet traffic across different times of the day (Beitzel, Jensen, Chowdhury, Grossman, & Frieder, 2004; Jansen et al., 2011). Beitzel et al. (2004) indicated a big change search queries’ frequency distribution in a day. Paul et al. (2011a) later showed the influence of posting time on response rate, given that tweets posted during peak hours could easily get buried in content streams. This leads me to investigate whether or not picking the right time to ask one question is important to getting a high response rate. I divided the posting timestamp of the collected questions into four categories of equal durations, including: nights (0:01AM - 6:00AM), morning (06:01AM – 12:00PM), afternoon (12:01PM – 6:00PM) and evening (6:01PM – 12:00AM), as inspired by findings from previous studies (Demirbas, Bayir, Akeora, Yilmaz, & Ferhatosmanoglu, 2010; Wakamiya, Lee, & Sumiya, 2011). As a categorical predictor, posting time period was dummy coded for later analysis.

5. **Topical Category**: Given that expertise is usually context dependent, I assume that the response rate of questions might be also distinct across topical categories. Focusing on the topical categorization of interrogative tweets, prior studies (Efron & Winget, 2010; Z. Liu & Jansen, 2012) presented significant differences on the number of questions posted across categories. Another study (J. Yang & Wei, 2009) showed such topical variance in people’s knowledge sharing behavior, indicating that certain categories tend to attract more answers than the others.
To examine this assumption, I employed a categorization method by automatically submitting each of the collected questions to Baidu Zhidao (http://zhidao.baidu.com/) and retrieving the returned classifications. As the most famous professional Q&A site, Baidu Zhidao contains 14 main categories, including: Computer and Internet, Life, Health, Sports, Electronics, Business, Education and Science, Society, Culture and Arts, Game, Entertainment, Personal Vexation, and Region. Under each main category, there are also a number of sub-categories. The most frequently occurring main category of the top five returned results would be assigned to the question as its topic. For questions containing more than five clauses, I shortened them using the sub sentence containing the question keyword with one clause before and one clause after.

6. **At-mentions and Hashtags:** The at-mention feature of Weibo enables users to directly reference others by putting an @ symbol before their screen names. According to Huberman, Romero, and Wu (2008), this feature is widely adopted by Twitter users with about 25.4% of all daily tweets being directed ones. At-mention is a strong predictor of information diffusion (J. Yang & Counts, 2010), as well as a significant factor in enlarging a post’s visibility and helping initiate responses and conversations (Vega, Parthasarathy, & Torres, 2010). The effect’s presence was shown in both Comarela, Crovella, Almeida, and Benevenuto (2012) and de Souza, de Magalhães, de Costa, and Fechine (2012)’s studies.

Hashtagging is the way Weibo and Twitter categorize posts according to specific keywords or topics; hashtags’ abilities to group conversations and information diffusion has been well studied. Rossi and Magnani (2012) investigated hashtag-based conversations and found that hashtags help break through the social network’s restricting structure and make conversations based on non-reciprocal following relationships possible. Based on their findings, I suppose users can enlarge the visibility of their questions by adopting hashtags, and thus having an increased possibility of getting responses.
7. **Emoticons:** Emoticons are graphic representations of facial expressions that Weibo users can embed in their post. Previous literature (Derks, Bos, & Von Grumbkow, 2007; Lo, 2008; Walther & D’Addario, 2001) suggested that in compensation of the lack of nonverbal cues, people tend to use more emoticons in computer-mediated communications as such enables them to better maintain their social presence and to be more engaged in social interactions. Besides, emoticons can also be used to draw attention from the recipients (Kriplean, Toomim, Morgan, Borning, & Ko, 2012). I expect a positive correlation between the emoticon use and question response rate.

8. **Question uniqueness:** This feature measures one question’s uniqueness as compared to others collected in the data set. Previous studies suggested that the amount of unique information in a question positive impacted the question’s quality (Kitzie, Choi, & Shah, 2013). They claimed that the unique information presented may help an answerer interpret the questioner’s information needs with a higher level of specificity, thus improving the answer’s overall quality. Additionally, question uniqueness has also helped in identifying high-quality social media content (Agichtein et al., 2008). Inspired by both studies, I adopted two metrics to measure the amount of unique information in a question posted on Weibo. First, I measured the percentage of words in a question that were not in the top 10 most frequent words in the collection. Second, I calculated the fraction of unique words in a question. For both measurements, both stop words and search keywords were removed before calculation.

**User Features**

The user features describe the author of a question via his/her profile characteristics, as well as past behaviors. Below are the user features adopted into this study.
1. **Number of Followers:** The number of followers has been investigated as a possible indicator of a user’s influence in spreading information to effective readers (Cha, Haddadi, Benevenuto, & Gummadi, 2010). Given that questions posted on one’s SNS can often only be seen by one’s followers/friends, I assumed that the more followers a questioner has, the more responses he/she could expect to receive. Consistent with my assumption, Paul et al. (2011a) concluded that the probability of receiving a response is intrinsically associated with the questioner’s number of followers. de Souza et al. (2012) also supported the relationship through their computational investigation.

2. **Posting Frequency:** Besides a large number of followers, a high posting frequency can also, to some extent, indicate an online influencer (Reynolds et al., 2010). Thus, it might have positive impact on the response rate of social Q&A as well; however, based on my own experience and past literatures, posting too much mundane information on one’s social network may cause negative physiological consequences to audiences and lead to information fatigue (Gelter, 2009). People may become uninterested in reading the messages posted by followees who disclose too much information (Oulasvirta, Lehtonen, Kurvinen, & Raento, 2010). Both Sibona and Walczak (2011) and Kwak, Lee, Park, and Moon (2010) demonstrated that people “unfriend” those who post too frequently about unimportant topics or mundane personal details. To calculate one’s posting frequency, I divided the total number of posts by the total number of days on Weibo from the first day of registration.

3. **Posting Style:** Every user has his/her own style of posting on SNS. Some prefer to retweet, while others like to engage in dialogues. Wallsten (2008) suggested that users can be grouped into different types based on the way they post on blogs. He showed that blogs were complex platforms that contain a mix of opinion statements, mobilization attempts, requests for audience feedback, and links to information produced by others. Z. Liu and Weber (2014) further indicated that the way people tweet reflected their behavior toward ideological friends and foes.
Inspired by Wallsten (2008), I measured one’s posting style on Weibo from four different dimensions, including their rate of retweet, at-mention, URL, and sharing.

**Social Activity Features**

The features of social activities captured the social attractiveness of the questioner, and his/her social engagement with followers based on past activities. I introduced three social activity features: the average number of retweets / comments / likes received: These features reveal the content value of the questioner, as well as his / her popularity among followers.

**Data Analysis**

To predict the response rate, I used logistic regression to calculate the probability of a new question being answered. A logistic regression is a form of regression equation where the dependent variable is dichotomous rather than continuous and the output is transformed into a probability (Stockwell, 1999) as described by the following equation:

\[
P(\text{response}_j | f) = \frac{1}{1 + e^{-(w_0 + \sum w_i f_{ij})}}
\]

I aim to calculate the response probability of question \( j \), given response \( j \) as the dependent variable of possible values of 0 or 1, and a set of independent features \( f_1, f_2, \ldots, f_k \), where \( f_{ij} \) be the feature \( i \) of question \( j \), and \( w_0 \) and \( w_i \) as the weights learned by the logistic regression model. If \( w_i \) is positive, then an increase in \( f_i \) is associated with an increase in response probability and vice versa.

Before I conducted the regression analysis, multicollinearity was checked by examining the bivariate correlations across all predictor variables. As none of the values in the bivariate
correlation matrix exceeded the recommended value of 0.7 (Slinker & Glantz, 1985), the data suggested a lack of multicollinearity among independent variables in the multiple linear regression models. Given the skewed distribution of all the numerical variables, I used a log-transformation to normalize the data. All categorical predictor variables were dummy coded. I employed SPSS for the analysis. The p-value was set at 0.05 to be statistically significant.

Results

Descriptive Statistics

Overall, the 62,106 information-seeking posts collected generated 607,497 responses. At least one answer was provided to 61.68% of the Weibo questions. I noticed that the response rate of my data set was relatively higher than a previous study (Paul et al., 2011a). I thought this might due to the keyword-based method adopted for question collection. While using keywords to retrieve questions, I could be excluding many conversational questions with relatively low response rates. In general, the response rate of Weibo questions followed a long tail distribution as shown in Figure 4-2.
Figure 4-2. Distribution of responses to questions posted on Weibo illustrating a long tail distribution.

Data Analysis

To answer the first research question, I modeled the question response rate using a logistic regression implemented in SPSS. The results are summarized in Table 4-1 with the coefficient, odds ratio, and the p-value associated with each feature.
Table 4-1. Results of logistic regression with coefficient, odds ratio, and p-value.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Coefficient</th>
<th>Odds Ratio</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Retweet</td>
<td>1.86</td>
<td>6.41</td>
<td>0.00***</td>
</tr>
<tr>
<td>% Being liked</td>
<td>0.90</td>
<td>2.45</td>
<td>0.00***</td>
</tr>
<tr>
<td>% Being retweeted</td>
<td>1.01</td>
<td>2.74</td>
<td>0.00***</td>
</tr>
<tr>
<td>% Being commented</td>
<td>5.70</td>
<td>228.22</td>
<td>0.00***</td>
</tr>
<tr>
<td># Followers</td>
<td>0.01</td>
<td>1.00</td>
<td>0.00***</td>
</tr>
<tr>
<td># Post per day</td>
<td>0.00</td>
<td>1.00</td>
<td>0.48</td>
</tr>
<tr>
<td>Question length</td>
<td>0.00</td>
<td>1.00</td>
<td>0.12</td>
</tr>
<tr>
<td>Repeated punctuations (binary)</td>
<td>0.00</td>
<td>1.00</td>
<td>0.94</td>
</tr>
<tr>
<td>Repeated Interjections (binary)</td>
<td>0.05</td>
<td>1.01</td>
<td>0.12</td>
</tr>
<tr>
<td>Gratitude (binary)</td>
<td>0.08</td>
<td>1.08</td>
<td>0.16</td>
</tr>
<tr>
<td>Emoticon (binary)</td>
<td>0.12</td>
<td>1.13</td>
<td>0.00***</td>
</tr>
<tr>
<td>Mention (binary)</td>
<td>0.63</td>
<td>1.98</td>
<td>0.00***</td>
</tr>
<tr>
<td>Hashtag (binary)</td>
<td>-0.38</td>
<td>0.69</td>
<td>0.00***</td>
</tr>
<tr>
<td>% words not in top 10</td>
<td>0.37</td>
<td>1.45</td>
<td>0.00***</td>
</tr>
<tr>
<td>% unique words</td>
<td>0.23</td>
<td>1.26</td>
<td>0.04*</td>
</tr>
<tr>
<td>Topical Category</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthcare-Entertainment</td>
<td>0.36</td>
<td>1.44</td>
<td>0.00***</td>
</tr>
<tr>
<td>Business-Entertainment</td>
<td>-0.11</td>
<td>0.90</td>
<td>0.19</td>
</tr>
<tr>
<td>Sports-Entertainment</td>
<td>-0.07</td>
<td>0.93</td>
<td>0.36</td>
</tr>
<tr>
<td>Education-Entertainment</td>
<td>-0.11</td>
<td>0.90</td>
<td>0.02</td>
</tr>
<tr>
<td>Culture-Entertainment</td>
<td>-0.01</td>
<td>1.00</td>
<td>0.86</td>
</tr>
<tr>
<td>Vexation-Entertainment</td>
<td>-0.23</td>
<td>0.80</td>
<td>0.00***</td>
</tr>
<tr>
<td>Life-Entertainment</td>
<td>0.17</td>
<td>1.18</td>
<td>0.00***</td>
</tr>
<tr>
<td>Technology-Entertainment</td>
<td>-0.02</td>
<td>0.98</td>
<td>0.49</td>
</tr>
<tr>
<td>Society-Entertainment</td>
<td>0.09</td>
<td>1.10</td>
<td>0.13</td>
</tr>
<tr>
<td>Time</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0:00-06:00AM - 18:01-23:59PM</td>
<td>0.02</td>
<td>1.02</td>
<td>0.69</td>
</tr>
<tr>
<td>06:01-12:00PM - 18:01-23:59PM</td>
<td>-0.36</td>
<td>0.70</td>
<td>0.00***</td>
</tr>
<tr>
<td>12:01-18:00PM - 18:01-23:59PM</td>
<td>-0.22</td>
<td>0.81</td>
<td>0.00***</td>
</tr>
</tbody>
</table>

Adjusted R2 = 0.40

*** p-value < 0.001, ** p-value < 0.01, * p-value < 0.05

From Table 4-1, I noticed that all proposed features, taken together, accounted for roughly 40% of the total variance. Although 40% might seem to be low, given that all 17 features studied in this work are extrinsic factors that may have impact on people’s social Q&A behavior without inclusion of any intrinsic determinant, the relatively lower R-square value in the model is reasonable.
In addition to an overall assessment of the prediction model, I also evaluated the proposed features by examining the p-values, coefficients and odds ratios associated with each of the predictor variables. I found no statistical evidence to support my hypotheses on posting frequency, question length, usage of repeated punctuations, and gratitude.

Findings from the multivariate logistic regression provided a general indication of the associations between hypothesized predictors and the question response rate; however, in order to better understand those associations, I conducted more in-depth analyses for each individual feature. Relevant patterns were documented and are discussed in the following section.

**Feature Analysis**

**Question Features**

**Emoticons**

While analyzing emoticons usage in social Q&A, I found that among the 24,898 questions with emoticons adopted, 16,510 (66.31%) received at least one answer; however, only 58.53% of the questions without the adoption of emoticons got responded. In other words, the adoption of emoticons did attracted more attentions and responses in social Q&A. In Table 4-2, I listed the top 10 emoticons adopted in questions received at least one answer. While grouping the top 10 emoticons into three sub-categories according to their underlying intent, namely, the *frustration* emoticons, the *appreciation* emoticons, and the *other* emoticons, I noticed that half of the most adopted emoticons were used to express frustration or sadness in not receiving adequate or timely support. In addition, I also found that compared with adopting frustrating emoticons,
using appreciation emoticons significantly increased the response probability in social Q&A, even though such gratitude was not clearly identified at the textual level.

Table 4-2. Top 10 emoticons adopted in responded questions.

<table>
<thead>
<tr>
<th>Emoticon</th>
<th>Frequency / Percentage</th>
<th>Responded Questions (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frustration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>😡</td>
<td>5,938 / 23.85</td>
<td>65.08</td>
</tr>
<tr>
<td>😠</td>
<td>2,162 / 8.68</td>
<td>61.74</td>
</tr>
<tr>
<td>😞</td>
<td>1,374 / 5.52</td>
<td>64.63</td>
</tr>
<tr>
<td>😥</td>
<td>1,009 / 4.05</td>
<td>65.54</td>
</tr>
<tr>
<td>🙁</td>
<td>808 / 3.25</td>
<td>65.58</td>
</tr>
<tr>
<td>Appreciation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>😊</td>
<td>1,470 / 5.90</td>
<td>71.89</td>
</tr>
<tr>
<td>😊</td>
<td>997 / 4.00</td>
<td>77.43</td>
</tr>
<tr>
<td>Others</td>
<td></td>
<td></td>
</tr>
<tr>
<td>😢</td>
<td>2,219 / 8.91</td>
<td>64.80</td>
</tr>
<tr>
<td>😓</td>
<td>1,775 / 7.13</td>
<td>70.00</td>
</tr>
<tr>
<td>😞</td>
<td>676 / 2.72</td>
<td>66.41</td>
</tr>
</tbody>
</table>

Mentions

A cross-tabular table was constructed as shown in Table 4-3. 76.28% of questions with @mentions were answered by at least one person, whereas only 59.13% of questions without @mentions got their reply. This indicated that mentioning a specific user in a question increased the probability of receiving a response.

Table 4-3. Cross-tabular distribution of @mention and responded type.
Hashtags

To my surprise, results shown in Table 4-1 indicated that the adoption of hashtag might actually decrease the probability of receiving an answer. While through a more detailed analysis I found that given no set rules for hashtag creation, 95.62% of all the hashtags adopted in the collected questions appeared only once or twice. Among all 5,128 questions with hashtag adopted, only 53.27% received an answer; whereas, 62.40% of the questions without hashtag got responded to by others. This indicates the hashtag’s negative impact on response rate, which is consistent with previous findings (Z. Liu & Jansen, 2013). Given hashtags’ heavy usage in spam posts (Walther & D’Addario, 2001), I surmised the negative impact of hashtag usage on question response rate might be because using hashtags can annoy readers and lead them to skip those.

Question Uniqueness

To further understand the effect of question uniqueness on question response probability, I plotted the diagram of cumulative distribution function for both measurements, as shown in Figure 4-3. The plot on the left shows the cumulative probability distribution of the fraction of words in a question not in the top 10 frequent words in the collection. I found that responded questions contained a higher amount of rare words compared to non-responded questions. Additionally, from the plot on the right side of Figure 4-3, I noticed that again responded questions contained less unique words compared to non-responded questions. This observation corresponded with my intuition that responded questions are more specified and personalized, so unique questions may assist the answerers in interpreting the questioner’s information needs, thus improving the response’s probability.
Figure 4-3. Cumulative distribution function of two question uniqueness measurements.

**Topics**

As shown in Figure 4-4, the number of questions asked by Weibo users varied largely among all 10 topical categories. The most popular topic on Weibo was Entertainment (15,166, 24.42%). This is consistent with the entertaining nature of SNS. In contrast, the least popular topic was Business (916, 1.48%).

From Figure 4-4, I further indicated that although the number of questions asked varied significantly across topical categories, the question response rate did not show huge differences, apart from a few exceptions. I found that questions regarding personal vexation had the lowest response probability, perhaps due to the ambiguity in user’s information-seeking intents. In contrast, questions in the topical categories of both “Healthcare” and “Life” received the highest chances of being answered in the data set with their response probability as 0.69 and 0.68.
respectively. Using Chi-square tests, such topical difference was proved to be significant for the question response rate at p < 0.01 ($\chi^2 = 945.89, p = 0.00$).

Figure 4-4. Distribution of topical categories of questions and their response probability.

Posting Time

I conducted an analysis on the temporal effect on social Q&A behaviors, as prior work as shown a time element to online interactions (Ying Zhang, Jansen, & Spink, 2009). As shown in Figure 4-5, I noticed that people asked the most questions between 18:01 PM and 23:59 PM (48.17%) and the fewest questions from midnight (3.15%) to the early morning (13.36%), which
is consistent with the officially reported pattern of Weibo usage from Sina. Quite different from the distribution of question posting, the response probability cross temporal categories demonstrated completely different distributions with questions posted from 0:00AM to 6:00AM (66.96%) and 6:01AM-12:00PM (67.12%) having the highest probability of being answered. Again, chi-square tests validated the statistically significant difference on question response rate across temporal periods at the level of $p < 0.01$ ($\chi^2 = 230.525$, $p = 0.00$). Such distribution difference suggested a gap between question asking and answering, indicating that although people are more active in information seeking from 12:01PM – 23:59PM, they actually have lower chances of answering another’s questions.

![Figure 4-5. Distribution of posting time period of questions and their response probability.](image)
Social Activities and Posting Style Features

According to the odds ratio from Table 4-1, I found that among all proposed predictors, the social activity based features have the most discriminative power on predicting question response rate. To further indicate such differences between responded to and non-responded to questions, in Figure 4-6 I deployed box plots for past social activities-based features as well as posting style-based one. As indicated by both the correlation coefficient in the logistic regression model and the box plot, I noticed that the more social interactions (like, retweet, comment) an individual received in the past, the higher probability he/she would get a response when asking a question on Weibo. Interestingly, I found that individuals who like to retweet have higher chances of getting their questions answered.

Figure 4-6. Box plots depicting the distribution for responded and non-responded social Q&A questions based on retweeting behaviors and past social activities.
Prediction Experiment

In addition to looking into each of the individual features, I also assessed the predictive power of the whole model to answer the second research question. I conducted a prediction experiment using the same features as proposed in the aforementioned logistic regression. The model was trained and tested using the algorithm of logistic regression as implemented in Weka (Hall et al., 2009) with 10-fold cross-validation. I balanced the data set to achieve an equal number of positive and negative instances. For evaluation purpose, I used the traditional metrics, including: precision, recall, F1 and accuracy, as they have been adopted in many other studies. Majority induction was adopted as the baseline model to interpret the classification results. With the balanced data set, I got a baseline accuracy of 0.50.

Table 4-4 lists the classification results of the proposed logistic regression model using all features and features only features from the user’s perspective, as well as from the question’s perspective. One can see from the table that the model performed the best when adopting all features from both the question and the user’s perspectives, achieved a classification accuracy of 0.74, which is much higher than the 0.5 baseline, and validated the quality of the proposed prediction model.

Interestingly, I noticed that adopting features from the user’s perspective alone achieved a prediction accuracy of 0.73, which is almost the same as the model with all features included. The results indicated that whether or not a question would be answered on Weibo was related more to the questioner than to the question’s content. This leads us to conclude the different nature of questions asked on SNS and community Q&A sites with question posted on SNS required less expertise. In addition, consistent with my analysis results as shown in Table 4-4, with a further analysis on the predictive power of the features within the user’s aspect, I found that features based on user’s past social activities contributed the most to the discrimination of the
responded and non-responded classes. This finding demonstrated the important role that social activities played in response rate prediction. In other words, more interactive users tended to receive higher probabilities to have their questions answered.

<table>
<thead>
<tr>
<th>Features</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question Features</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>User Features</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>Social Activity Features</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>Question + User Features</td>
<td>0.74</td>
<td>0.74</td>
<td>0.74</td>
<td>0.74</td>
</tr>
</tbody>
</table>

**Discussion and Implication**

My research work presented in this chapter is valuable to people in the field of social information seeking, ranging from social scientists to social network users. Given the lack of studies focusing on the response quantity of social Q&A, my predictive model can be adopted as the foundation for future studies with the related focus. Social technologists could also benefit from this study by understanding user’s behaviors toward various features when they decide to interact and help others by responding to questions. Taking advantage of my findings and suggestions can help social technologists know the bottlenecks of current systems in the social Q&A environment so that they can develop new tools to support user’s questioning and answering behaviors on SNS. When asking questions on SNS, people can utilize the results of this study to match their conditions, checking to see if their question has the potential to receive an ideal number of replies. In summary, this work is of positive value to the research community. I aim to build tools, such as an expert recommender, or a question routing mechanism, to make the social Q&A process easy for users to use.
Limitations of this current work include the keyword method of extracting interrogative posts. Although those questions are representative given their high frequency of occurrence, only focusing on this method of extraction might affect the generalizability of work, especially when facing uncommonly asked questions. As discussed earlier, with only extrinsic features considered, the current model may not achieve high prediction. Adding more intrinsic and social predictors into the model will increase its predictability. The investigations on the relationships between social predictors and the response behavior could be one possible direction for later work.

**Conclusion**

With the objective to enhance user interactions and cooperation in a social Q&A context, I performed analyzed the relationships between extrinsic features and question response rates. Using Weibo as data source and logistic regressions as statistical method, findings are that features from the perspectives of past social activities, posting behavior, and question phrasing significantly impact the response rate. The adoption of hashtags negatively influences the chances of receiving answers.

Through further analysis of those predictors, I noted design problems from two perspectives that may impact the usability of social Q&A. Since there are no official rules for hashtag usage from major SNS providers, hashtags might not be beneficial in the context of social Q&A, given that they do not increase the visibility of the question. Also, the asynchronous nature in social Q&A has been demonstrated while analyzing the response probabilities across question posting time period.

Given the present situation, developing corresponding tools and mechanisms, such as interrogative hashtag vocabularies, question recommendation, and compensation could make the
user’s expert seeking process easier and make the experts more willing to share knowledge with others. Based on these possible implications proposed, I believe that my study offers valuable insights into the future development of social search systems or tools that can make good use of those features as introduced in this study.
Chapter 5

Identifying the Willingness to Response: Characterizing Knowledge Contribution of Individuals in Social Q&A

Due to the advantages of social Q&A as introduced in previous sections, it has attracted the attention of many researchers, and has motivated the creation of models and tools to facilitate the information seeking process. Among the proposed methods are varies questions routing algorithms, mostly involving expert finding techniques to address the problem of nonguaranteed responses under social context. Although routing algorithms based on expertise could effectively recognize individuals who can answer a specific question, it still suffers from the challenge of identifying people who are willing to do so. In other words, even after finding people with required knowledge, we still need a way to know if they are willing to respond to stranger’s questions.

To solve this issue, in this study I analyze a collection of questions and answers posted during a 10-month period on Weiwen, a Chinese question routing services based on microblogging sites. I explore the patterns demonstrated by the knowledge contributors in the social Q&A process from three different perspectives: user behaviors, user interest, and user connectedness. I also propose a predictive model on knowledge contributor in social Q&A environment based on a number of non-Q&A features. I found that less popular but more interactive individuals are more willing to respond to strangers in social Q&A environments. In addition, I also noticed that the psycho-linguistic characteristics of an individual’s SNS posts, such as their usage of verbs, pronouns, and cognitive expressions, can also indicate their roles in Q&A. I believe this research provide benefits by providing a more in-depth understanding of social Q&A, especially what are the characteristics of people voluntarily answering stranger’s
questions. It can also be viewed as a design guideline for future question routing systems based on SNS.

Background

As noted in prior research (Z. Liu & Jansen, 2013; Paul et al., 2011a), majority of questions posted on SNS received no response. In order to solve this problem, many recent studies have been conducted with their focus on routing questions to the appropriate respondents for help. Through online experiments, (Nichols & Kang, 2012) explored the feasibility of users responding to questions sent by strangers. They found that less than half of the people will answer to questions posted by strangers; however, they failed to indicate the characteristics of those responders. Pan et al. (2013) offered a more in-depth analysis on potential answerers by leveraging users’ non-Q&A social activities. Through their analysis of an inter-organizational community Q&A site, they found that some of the non-Q&A features can effectively predict the likelihood of one’s answering of others’ questions. Luo et al. (2014) built a Smart Social QA system based on IBM Connected that recommends both active and inactive users for a given question based on their based on their abilities, willingness, and readiness. The only limitation of their work is that given the differences between organizational community Q&A site and SNS, their framework may meet with some difficulties when generalized to social Q&A context.

Considering the lack of study and the importance of identifying potential responders in social Q&A, I conducted this work to understand what kinds of users are more willing to respond to stranger’s questions on SNS. Given the very response rate reported in Nichols and Kang (2012)’s study, I believe that without a more comprehensive evaluation of characteristics signifying a knowledge contributor in social Q&A, SNS’s power in social information seeking may never be achieved.
Research Objectives

The ultimate goal of this study is to improve the effectiveness of question answering that happen on SNS by routing questions to individuals with desire to help others. In order to achieve this goal, I present two specific research objectives:

1. Explore the question answering patterns of individuals when they are exposed to questions asked by strangers via question routing.

For this research question, I analyze the question answering data collected from a Chinese social Q&A application called Weiwen (微问). Especially I am interested in characterizing individual’s answering behaviors, their response interests, and their social connectedness in the question answering process. By answering the first research question, we are able to evaluate the effectiveness of question routing in social Q&A.

2. Identify individuals with the desire to help others in social Q&A by using their non-Q&A characteristics.

The second research question aims to explore the distinction between active users and non-active users in reacting to stranger’s questions routed to them. Here by saying “active” users, I mean individuals with high willingness to answer other’s questions. To measure the differences, I introduced non-Q&A features from three different aspects, such as individual’s SNS profile, and posting behaviors. QA-related attributes such as the difficulty of the question asked, and the knowledge and expertise level of the answerer, are not considered in this study. In achieving the
second research objective, I am able to contribute to the design of future question routing services or tools, which aims at directing questions to individuals who are likely to provide answers.

Weiwen and Data Set

To understand the knowledge sharing behaviors in social context, I collected data from a social Q&A application based on Sina Weibo, called Weiwen (微问, http://weiwen.weibo.com/). Weiwen operates in a different manner compared to traditional community Q&A sites, such as Yahoo! Answers and Baidu Knows. In those traditional community Q&A sites, people ask questions and then passively wait for the potential helpers to see and respond. In contrast, in Weiwen, individuals can either post questions directly to the site, or can post their questions on Weibo by mentioning @微问 (@Weiwen). After receiving the questions, Weiwen will next identify a number of potential respondents based on their expertise and experience as demonstrated on their Weibo profiles, using machine learning techniques. By routing questions to those “qualified” respondent, Weiwen effectively increases the probability of obtaining high quality response. A graphical demonstration of the question routing procedure of Weiwen is shown in Figure 5-1.
Another difference between Weiwen and other community Q&A sites is that, in addition to presenting the answers received, in most cases, Weiwen also informs the users to whom the question has been routed. This allowed us to know who has responded a question and who hasn’t, and it enabled us to build the classifier with both positive and negative instances. Figure 5-2 is a screenshot of Weiwen, major sections highlighted.

With Weiwen, I crawled in total 340,658 questions from January 24th, 2013 to October 18th, 2013, together with 1,754,280 replies, and 585,359 unanswered records.
Figure 5-2. Screenshot of Weiwen with the major sections of Question Routed, Answers Received, and Non-responders highlight, along with the Question Category.

Results Characterizing Knowledge Sharing Behavior on Weiwen

Aggregate Analysis of Question and Answering

Based on an initial examination of the data set, I noticed that 339,878 out of all 340,658 questions in the collection received at least one answer, yielding a response rate of 99.77%. On average each question received 5.14 answers. Comparing with the relative low number of answers received in natural social Q&A settings, the data set revealed the promising performance of question routing in a real social Q&A services.

In order to examine the patterns of knowledge exchange in social Q&A, I further analyzed the roles that individuals played in Weiwen. I found that in total 671,501 Weibo users
participated in the social Q&A process. Among them, 22,203 (3.31%) individuals both asked and answered questions. 221,060 (32.92%) asked at least one question, but provided no answer. In contrast, 472,644 (70.39%) users posted no question, but replied at least once on Weiwen. The 340,658 questions were asked by 243,263 unique individuals and answered by 494,847 ones. In Figure 3, I plotted the distributions of the number of questions asked and the number of answers provided by each Weiwen user collected in the data set. Surprisingly, I noticed that there were more contributors (users who posted more answers than questions) than consumers (users who posted more questions than answers) on Weiwen. Considering the large number of contributors, the high response rate on Weiwen is not surprising. Besides, while comparing my results on contributors versus consumers with the findings presented in Shah et al. (2008) and Gyongyi, Koutrika, Pedersen, and Garcia-Molina (2008) observations based on Yahoo!Answers, again, I noticed the power of question routing in social Q&A context. I plotted the cumulative probability distribution of the total number of questions answered on a log-log scale and noticed that the number of questions answered on Weiwen followed a power law distribution, as the points fell closely on the straight line in the log-log plot. This indicated an uneven participation in Weiwen, where a small number of individuals contributed to a large proportion of questions, and a large proportion of users only answered a few number of questions.
In addition to the analysis on the questioning and answering behaviors, I also explored the topical interests of individuals in Weiwen. To better organize questions according to user’s interests, Weiwen grouped all questions posted on it according to a topical hierarchy containing 13 top level categories, including “Computer / Network”, “Game”, “Sports”, “Healthcare”, “Life”, “Vexation”, “Business”, “Education / Science”, “Society”, “Culture / Arts”, “Digital Electronics”, “Entertainment”, and “Resource Sharing”. Under each top category, there are a number of sub-categories.

I plotted the number of questions belonging to each topical category in a bar diagram in Figure 5-4 (a). I observed that the topical category of “Life” contained the highest number of
questions (32.05%), followed by the categories of “Entertainment” (27.72%). These two categories account for more than half of the questions asked on Weiwen, with the remainder 40% of questions distributed among the other 11 categories. While examine the average number of responses received across categories, I noticed from Figure 5-4 (b) that questions under the topics of “Entertainment”, “Vexation”, “Life”, “Electronics”, and “Arts” obtained the highest number of answers on average. I thought this might be due to the subjective nature of questions under these two topical categories, as well as the low expertise required to answer them. Example questions with higher number of responses from the abovementioned categories are: “Can anybody recommend any horror movie for me please?”, “Has anybody used the Sony nex3n yet? Thoughts?”, “Do girls really care about height that much? Like i'm 175cm so it's minor, just curious?”, etc. On the other side, questions under the topical categories of “Sports”, “Education”, and “Computer” received lower number of answers, indicating their relatively objective nature. Some typical factual-seeking questions under those categories include: “When should I register for the GRE test?”, “So how can i get to the reach of root of locked iphone?”

![Figure 5-4](image-url)

Figure 5-4. Topical distribution of questions asked on Weiwen by percentage of questions posted by topic, number of answers received by topic, and word length of both questions and answers.
I next explored the ways individuals ask and answer questions on Weiwen. I assume that one could infer the type of information needs under each topic by examining the average question and answer length in words within that category. As can be seen from Figure 5-4 (c), I noticed that questions under the topical categories “Entertainment”, “Life”, “Electronics” and “Computer” were asked in a fairly specific manner, whereas were answered with many short replies. For example, the question “My IPAD suddenly go dead and after reboot all my downloads disappeared, and now I can't even download anymore! I get a warning message saying 'Make sure SD card is writeable', isn't IPAD SD card internal, what's going on?” attracted many general answers, such as “Reset to factory settings”, “Due to the loose connection of your SD card, get a replacement at Apple if within warranty”, “Better contact Apple support”, etc. In contrast, questions under the topical categories “Vexation” “Healthcare”, “Education”, and “Business” were phrased in a relative general manner, as for example the question “Why do my gums bleed when I brush my teeth?” received long replies such as “First rule out the possibility of blood system disease and weak liver and spleen. Then if your bleeding is caused by an accumulation of plaque around gums, ultrasonic cleaning may provide some effective relief”. “I would suggest to take a blood routine examination to check the common indices of coagulation. Most of times it is a symptom of gingivitis, pay attention to your oral health and the correct way of brushing your teeth, don’t eat spicy food.”

After examining the overall distribution of questions and answers across topics, I further looked at the contributions of each specific user in the data set. I found that the average number of topics in which a user post a question is 1.07, which is much lower than the average number of topics user responded (1.33). 85.23% of the contributors answered questions under only one single topic, whereas 62,160 answerers responded to questions from within more than 1 topical category, and these users generated more than 61.00% of all answers on Weiwen.
To further investigate the answerer distribution across topical categories, I plotted in Figure 5-5 the distribution of topics answered as a function of the total number of questions responded by each individual user in the data set. From the plot, I noticed that the number of topical categories answered by each individual increased along with the amount of questions they have in total responded. While analyzing the topical distribution of the top 1000 contributors, I noticed that on average they answered on average 503.13 questions within 12.22 topics. The Spearman correlation coefficient between the number of topics answered and the total number of answers provided per user is 0.71, indicating a strong association between individual’s responsiveness and diversity of interest.

![Figure 5-5](image)

**Figure 5-5.** Distribution of number of topics answered and questions answered, showing high correlation between responsiveness and interest diversity.

To unveil the association between topical categories, I measured the extent to which individuals answer questions in one category are also likely to do so in another. I adopted a method called Normalized Pointwise Mutual Information (NPMI) to estimate the collocation
strength between two different topics and expected findings, such as people who like to answer “Healthcare” rated questions are also likely to answer questions under the topical category of “Vexation”.

NPMI was proposed by Bouma (2009) as a bi-directional association measurement of the information overlap between two random variables, and it is a method that has been adopted to determine the degree of association between two events (de Souza et al., 2012; Yongzheng Zhang & Pennacchio, 2013). Compared to Pointwise Mutual Information (PMI) (Church & Hanks, 1990), the results of NPMI are easier to interpret and at the same time less sensitive to occurrence frequency. In my settings, for a pair of topics $T_x$ and $T_y$, I calculated their association by relating the probability of questions within $T_x$ and $T_y$ being answered by the same users with the probabilities of questions within $T_x$ and $T_y$ being answered individually. To be more specific, I estimate NPMI as follow:

$$NPMI(T_x, T_y) = \frac{\log \frac{P(T_x \cap T_y)}{P(T_x)P(T_y)}}{-\log(p(T_x, T_y))}$$

Equation 1: Formula to estimate NPMI

To transform the probability distributions into observable frequencies, I defined four variables $U_x$, $U_y$, $U_{xy}$, and $U$ denoting the number of users who have answered questions within the topical category $T_x$; the number of users who have answered questions within the topical category $T_y$; the number of users who have answered questions under both topics; and the total number of users within the data set. So the formula of NPMI can be rewritten as:
Equation 2: Revised formula to estimate NPMI

\[
\text{NPMI}(T_x, T_y) = \frac{\log \left( \frac{U_{xy}}{U_x U_y} \right)}{-\log \left( \frac{U_{xy}}{U} \right)}
\]

The result of NPMI ranges from \(-1\) to 1, with the positive value indicates the association of appearing together and the negative value indicates the association of not appearing together, 0 indicates statistically significant independent. I plotted the calculated NPMI value for each topic pair within the data set in Figure 5-6.

Figure 5-6. The NPMI values between topic pairs, with associations displayed in seven level scale from \((-0.3, -0.4)\) to \((0.9, 1.0)\), lowest to highest association.
As can be seen from Figure 5-6, I discovered that interestingly, individuals answering questions under the topical categories of “Life” and “Entertainment” tended to limit themselves to the current topics only, and would not answering questions under other topical categories. In contrast, I also noticed a number of topic pairs with positive associations, such as “Game” and “Sharing” (0.45), “Sharing” and “Vexation” (0.41), “Health” and “Vexation” (0.38), “Computer” and “Game” (0.38), etc.

**Network Analysis**

Next, in order to understand the structure of the Q&A networks formed on Weiwen, I applied the bow tie structure analysis (Broder et al., 2000) on the data set. The bow tie structure captures complex network structures. The key idea of the method is that a network can be viewed as a bow tie that is connected with four different components: Core, In, Out, and Tendrils/Tube, as shown in Figure 5-7. The bow tie structure analysis has been used in previous studies analyzing the network structure of Yahoo!Answers (Lin Chen & Nayak, 2011; J. Zhang, Ackerman, & Adamic, 2007).

In order to fit the collected Weiwen data into the bow tie model, I created a questioner-answerer graph by connecting users who asked questions with users who responded to these questions. Each node within the graph represents a user who asked or answered a question, while each edge corresponded to the directed reply relationship between the questioner and the answerer. The graph contained in total 715,209 nodes and 1,538,427 edges. The CORE component is the largest strongly connected component (SCC) of the questioner-answerer graph, in which any two users are mutually reachable by following the direct question-answering relationship. With the core component, I can detect the largest group of individuals who tend to help each other directly or indirectly on Weiwen. The IN component contains all nodes that are
not part of the CORE, but can reach it via directed paths. Users who always ask questions but rarely answer will primarily belong to the IN component. Similarly, the OUT component contains nodes that are reachable from the CORE via directed paths and in my case represents users who answer but infrequently ask. The “Tendrils” and “Tube” component (T&T) contains users who ask or answer only questions posted or responded by the users within the “IN” and “OUT” components.

Figure 5-7. The Bow Tie Structure (Broder et al., 2000)

For the QA graph generated using collected data set, the CORE component contains 9,552 nodes, which corresponded to 1.33% of all the users, which is quite different from the results as reported in previous studies on Yahoo!Answers. This indicated that the question answering process on Weiwen is not as social as one expects. Only a small proportion of users are connected on Weiwen through question and answering activities, while most of the users are quite segregated. In addition, to evaluate the reciprocal relationships between the questioner and the answerers, I also counted the number of mutual edges in the created graph. I found that
among all 1,538,427 edges, 9,313 (0.60%) were mutually connected. I believe that this indicated the well-separated roles played by the “contributors” and “consumers” in social Q&A environments, like Weiwen.

To test whether user connectedness correlate with topical interest, I also measured the size of the largest SCC and the number of mutual edges within all 13 topical categories. Table 5-1 shows the results obtained. Compared with the SCC measurement as reported in Adamic et al. (2008)’s study, I noticed from the table that the percentage of the nodes within the CORE for each individual category is much smaller that for the whole data set. None of the topical categories in Weiwen were well connected, as all largest SCC contained less than 1% of the users within that topical category. I thought this might be due to the well-separated roles of contributors and consumers on Weiwen, as I discussed in earlier section.

<table>
<thead>
<tr>
<th>Topical Category</th>
<th># Nodes</th>
<th># Edges</th>
<th>CORE (%)</th>
<th>IN (%)</th>
<th>OUT (%)</th>
<th>T&amp;T (%)</th>
<th>DISCONN (%)</th>
<th>MUTUAL EDGES (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Life</td>
<td>242,707</td>
<td>472,664</td>
<td>0.52</td>
<td>23.03</td>
<td>6.00</td>
<td>56.58</td>
<td>13.87</td>
<td>0.21</td>
</tr>
<tr>
<td>Entertainment</td>
<td>294,473</td>
<td>462,439</td>
<td>0.32</td>
<td>20.02</td>
<td>7.15</td>
<td>60.02</td>
<td>12.48</td>
<td>0.09</td>
</tr>
<tr>
<td>Healthcare</td>
<td>49,923</td>
<td>106,711</td>
<td>0.66</td>
<td>30.37</td>
<td>1.52</td>
<td>62.70</td>
<td>4.75</td>
<td>0.40</td>
</tr>
<tr>
<td>Arts</td>
<td>27,790</td>
<td>40,516</td>
<td>0.03</td>
<td>2.43</td>
<td>0.29</td>
<td>39.43</td>
<td>57.82</td>
<td>0.16</td>
</tr>
<tr>
<td>Sports</td>
<td>19,719</td>
<td>22,948</td>
<td>0.02</td>
<td>1.54</td>
<td>0.10</td>
<td>63.16</td>
<td>35.18</td>
<td>0.07</td>
</tr>
<tr>
<td>Society</td>
<td>25,820</td>
<td>41,869</td>
<td>0.30</td>
<td>11.63</td>
<td>1.14</td>
<td>90.26</td>
<td>3.33</td>
<td>0.27</td>
</tr>
<tr>
<td>Business</td>
<td>24,024</td>
<td>40,042</td>
<td>0.22</td>
<td>14.01</td>
<td>0.76</td>
<td>74.34</td>
<td>10.89</td>
<td>0.17</td>
</tr>
<tr>
<td>Electronics</td>
<td>41,708</td>
<td>88,703</td>
<td>0.34</td>
<td>16.90</td>
<td>1.18</td>
<td>75.57</td>
<td>6.01</td>
<td>0.19</td>
</tr>
<tr>
<td>Education</td>
<td>49,514</td>
<td>73,637</td>
<td>0.23</td>
<td>9.21</td>
<td>1.11</td>
<td>72.68</td>
<td>17.01</td>
<td>0.16</td>
</tr>
<tr>
<td>Computer</td>
<td>45,098</td>
<td>77,479</td>
<td>0.35</td>
<td>13.79</td>
<td>1.09</td>
<td>71.80</td>
<td>12.97</td>
<td>0.19</td>
</tr>
<tr>
<td>Game</td>
<td>13,704</td>
<td>18,720</td>
<td>0.28</td>
<td>5.11</td>
<td>0.76</td>
<td>77.10</td>
<td>16.76</td>
<td>0.30</td>
</tr>
<tr>
<td>Vexation</td>
<td>19,359</td>
<td>32,639</td>
<td>0.59</td>
<td>12.12</td>
<td>1.88</td>
<td>73.73</td>
<td>11.68</td>
<td>0.44</td>
</tr>
<tr>
<td>Sharing</td>
<td>1,364</td>
<td>1,254</td>
<td>0.01</td>
<td>1.32</td>
<td>0.07</td>
<td>28.74</td>
<td>69.86</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Table 5-1. Connectedness statistics for questions within 13 topical categories in Weiwen.

A further look at the results in Table 5-1 reveal that individuals who posted and answered questions under the topical category of Health, Vexation and Life are relatively connected, with
both a high percent of users within the largest SCC and a relatively large number of mutual edges. In other words, as compared to other topics, users focusing on those three categories were more likely to answer each other’s questions. This indicated the existence of connected communities within the social Q&A process under those topics. I thought this might because of the common ground existed between individuals sharing either the same living background (as many of the Life questions were location specific, so only users from the same regions can answer those questions), or physical (individuals who experienced or known someone who had the diseases) or emotional conditions (individuals who had or known someone who suffered from the vexation). In contrast, users within the topical categories such as Arts, Sports, and Sharing, are relatively less connected.

Besides, I also noticed that among all topical categories, there are much higher number of users contained within the IN component than in the OUT component, especially for the topical categories of Health, Electronics, Computer, and Business. This is consistent with the nature of the site Weiwen, where users actively came to seek for help. However, compared with my previous results as shown in Figure 5-3, I noticed that the larger number of answers is provided by only a smaller number of active answerers.

Predicting Super Contributors using Non-Q&A Features

By far, I have examined the knowledge sharing behaviors among relative strangers in Weiwen. From my analysis, I noticed the importance of identifying active contributors in social Q&A environments, as they not only provided the largest proportion of answers, but they are also part of the largest component within the questioner-answerer graph. So in this section, I proposed an experiment on predicting the super contributors in social Q&A environments.
Many of the past studies focusing on this topic relied on user’s past Q&A records (Dror, Koren, Maarek, & Szpektor, 2011; Jurczyk & Agichtein, 2007; Wang et al., 2014), however, as suggested by Pan et al. (2013) answerer prediction relying only on historical Q&A records suffers from the risks of losing the potential contribution of those who are new to social Q&A services due to the cold-start problem. In order to be more effective in such cold-start recommendation conditions, in this study, I build a model based only on non-Q&A traces. I derived those non-Q&A features from one’s Weibo account as linked to their Weiwen profile. I assume that individual’s online social behaviors and patterns to some extent can reflect their internal traits, and thus can be used to predict their potentiality in answering other’s questions.

**Data Set**

In order to identify the active contributors, I picked all users who have answered more than 50 questions in the collected data set. Besides, given the unequal number of questions routed to those users, I further constrained my sample based on their response rate. Only individuals who have answered more than 90% of questions routed to them have been included in my sample. I also manually tagged for enterprise and marketing accounts, and removed them from the sample considering the advertising nature of both their answers and their Weibo posts. In this way, I collected 1,503 active contributors in the data set.

To form the negative samples of inactive contributors, I included individuals with response rates of less than 10%. In addition, to ensure that those users were repeatedly not answering questions being routed to them, but not by chance, I only selected users with more than 4 unanswered questions. Again, I removed the enterprise and marketing accounts in order to be consistent with the sample of the active contributors. This left us with a total of 2,980 (include...
percentage) inactive contributors, based on which I randomly selected equal number of instances to produce a balanced data set.

**Feature Engineering**

Considering the goal of my second research question is to predict active contributors in Weiwen based on no historical Q&A records, I in this study introduced three non-Q&A feature classes based on individual’s Weibo feeds, including: user profile, user posting behavior, psycholinguistic traits of user posts, and social activities. Next, I will introduce each of the four types of features in detail.

**Profile Features**

The profile features indicate the identity of users, as well as their activeness in virtual worlds. Features such as gender, number of followers, etc., have been investigated in a number of prior studies as indicators of individual’s intrinsic characteristics. For instance, Ross et al. (2009) found a positive correlation between extraversion and the number of Facebook friends. Correa, Hinsley, and De Zuniga (2010) noticed that for females, there existed positive correlations between their openness and extraversion. Inspired from both studies, the profile features that I adopted in my model include: gender, whether or not is a verified account, number of followers, number of followees, longevity of the account, and posting frequency per day. The first four features can be retrieved directly from one’s Weibo profile, while the posting frequency feature can be calculated by dividing the total number of status by the longevity of the account.

**Posting Behavior Features**
Posting behavior features capture the way individuals use social networking services and how they interact with others. Drawing on previous studies, I believe that every user has his/her own style of posting on SNS. Some prefer to retweet, while others like to engage in dialogues. Wallsten (2008) suggested that users can be grouped into different types based on the way they post on blogs. He showed that blogs were complex platforms that contain a mix of opinion statements, mobilization attempts, requests for audience feedback, and links to information produced by others. Java, Song, Finin, and Tseng (2007) suggested that users rarely post but with many followers tend to be information seekers, while users who often post URLs in their tweets are most likely to be information providers. To verify these claims, in this paper, I tested 10 posting behavior features, including: the percentage of posts that are retweets, the percentage of posts containing at-mentions, URLs, hashtags, and emoticons, the percentage of original posts containing images or videos, and the average length of the posts. To quantify each of these features, I collected up to 450 most recent posts for each active and non-active user in the data set.

**Psycho-linguistic Features**

Features under this category provide a comprehensive description of the psychological and linguistic characteristics of individuals on SNS, and in turn help in better classification of users based on their internal traits. To measure those psycho-linguistic features, Pennebaker, Francis, and Booth (2001) created a Linguistic Inquiry and Word Count program (LIWC), which maps the relative word frequency to a set of psychological dimensions, such as linguistic dimensions (e.g., pronouns, tense), psychological constructs (e.g., positive motion), and personal concerns (e.g., leisure, death). Given its popularity, LIWC has been widely adopted as a psycho-linguistic measurement in prior studies (J. Chen, Hsieh, Mahmud, & Nichols, 2014; Garimella, Weber, & Dal Cin, 2014; Tumasjan, Sprenger, Sandner, & Welpe, 2010).
To measure the linguistic features in Weibo posts I used a simplified Chinese version of LIWC called TextMind (Gao, Hao, Li, Gao, & Zhu, 2013). However, slightly different from LIWC, TextMind defines 71 word categories (68 in LIWC), each containing a number of corresponding words. With the help of TextMind, I counted the number of words used under each of the 71 categories for every active and non-active user in the data set. I normalized the counts of every categories with the total number of words appeared in all 450 posts.

**Social Activities Features**

The features of social activities capture the social attractiveness of the questioner, and his/her social engagement with followers based on past activities. I introduced in total 3 features of social activity: The average number of retweets / comments/ likes received: these features reveal the content value of the questioner, as well as his / her popularity among followers.

**Classification Algorithms and Evaluation Metrics**

With the above features, I next built a binary classifier to automatically differentiate active contributors from the non-active ones. I trained and tested my model using a number of classification algorithms implemented in Weka (Hall et al., 2009), including: NaïveBayes, SVM (SMO) and Logistic Regression, using 10-fold cross-validation. For evaluation purpose, I used the traditional metrics, including: precision, recall, F1 and accuracy, as they have also been adopted in many other studies (Castillo et al., 2011; Yandong Liu, Bian, & Agichtein, 2008). The majority induction algorithm, which simply predicts the majority class, was applied to determine the baseline performance of the classifier. Since I have balanced the data into 50% positive
samples versus 50% negative ones as mentioned in the data collection section, with this approach I get a baseline accuracy of 50%.

Classification Results

For all three classification algorithms, I found the logistic regression achieved the best performance. A summary of the results were shown in Table 5-2. The classifier using features from all four perspectives achieved a prediction accuracy of 70%, which is slightly higher than the baseline accuracy of 50%. Although a prediction accuracy of 70% is not high, considering only non-QA features have been adopted, I deemed this method suitable to solve the cold start problem in predicting potential contributors in social Q&A. The profile-based features showed very limited effect on identifying potential answerers. However, prediction based the psycholinguistic characteristics of one’s posts demonstrated the best discriminative power.

Table 5-2. Classification results using Naïve Bayes, SMO, and Logistic Regression.

<table>
<thead>
<tr>
<th>Features</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Profile Features</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.57</td>
<td>0.56</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td>SMO</td>
<td>0.58</td>
<td>0.57</td>
<td>0.57</td>
<td>0.56</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td><strong>Posting Behavior Features</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.60</td>
<td>0.59</td>
<td>0.59</td>
<td>0.59</td>
</tr>
<tr>
<td>SMO</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.62</td>
</tr>
<tr>
<td><strong>Psycho-linguistic Features</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>SMO</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.66</td>
<td>0.66</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td><strong>Social Activity Features</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.59</td>
<td>0.59</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td>SMO</td>
<td>0.62</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.61</td>
</tr>
</tbody>
</table>
Table 5-3 summarizes the results of logistic regression using features from all four categories. Given the large number of features included in the model, here I only listed significant features with their correlation coefficient and p-value. As one can see from the table, in general less popular and less social users on Weibo actually contributed more to answering stranger’s questions in social Q&A. Next, I will explain each of the significant features in more details.

Table 5-3. Results of Logistic regression with coefficients and p-values of each feature.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Coefficient</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Profile Features</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Num of Reciprocal Relations</td>
<td>0.03</td>
<td>0.00***</td>
</tr>
<tr>
<td>Num of Follower</td>
<td>-0.01</td>
<td>0.00***</td>
</tr>
<tr>
<td><strong>Posting Behavior Features</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent of Mention</td>
<td>1.83</td>
<td>0.00***</td>
</tr>
<tr>
<td>Percent of URL</td>
<td>0.89</td>
<td>0.01**</td>
</tr>
<tr>
<td>Percent of Retweet</td>
<td>-0.67</td>
<td>0.01**</td>
</tr>
<tr>
<td><strong>Psycho-linguistic Features</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pronouns (total pronouns, I, them, itself)</td>
<td>-111.96</td>
<td>0.02*</td>
</tr>
<tr>
<td>Personal Pronouns (I, them, her)</td>
<td>-618.83</td>
<td>0.00***</td>
</tr>
<tr>
<td>Second Person Plural Pronouns (you, honorific of you)</td>
<td>-440.02</td>
<td>0.00***</td>
</tr>
<tr>
<td>I (I, me, mine)</td>
<td>-516.26</td>
<td>0.00***</td>
</tr>
<tr>
<td>We (we, us, our)</td>
<td>-515.66</td>
<td>0.00***</td>
</tr>
<tr>
<td>You (you, your, thou)</td>
<td>-489.72</td>
<td>0.00***</td>
</tr>
<tr>
<td>Shehe (she, her, him)</td>
<td>-497.23</td>
<td>0.00***</td>
</tr>
<tr>
<td>They (they, their, they’d)</td>
<td>-516.42</td>
<td>0.00***</td>
</tr>
<tr>
<td>Impersonal Pronouns (it, it’s, those)</td>
<td>103.47</td>
<td>0.03*</td>
</tr>
<tr>
<td>Verb (common verbs)</td>
<td>11.10</td>
<td>0.01**</td>
</tr>
<tr>
<td>Filler (blah, I mean, you know)</td>
<td>93.3</td>
<td>0.00***</td>
</tr>
<tr>
<td>Insight (think, know, consider)</td>
<td>32.73</td>
<td>0.01**</td>
</tr>
<tr>
<td>Cause (because, effect, hence)</td>
<td>40.53</td>
<td>0.01**</td>
</tr>
<tr>
<td>Tentative (maybe, perhaps, guess)</td>
<td>28.33</td>
<td>0.01**</td>
</tr>
<tr>
<td>Feel (feels, touch)</td>
<td>50.32</td>
<td>0.01**</td>
</tr>
<tr>
<td>Feature</td>
<td>Coefficient</td>
<td>p-value</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>Work (job, majors)</td>
<td>-10.09</td>
<td>0.03*</td>
</tr>
<tr>
<td>Achieve (earn, hero, win)</td>
<td>-17.72</td>
<td>0.03*</td>
</tr>
</tbody>
</table>

**Social Activity Features**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Coefficient</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent of Liked Post</td>
<td>2.38</td>
<td>0.00***</td>
</tr>
<tr>
<td>Percent of Retweeted Post</td>
<td>-2.28</td>
<td>0.00***</td>
</tr>
<tr>
<td>Percent of Total Interactions</td>
<td>-1.55</td>
<td>0.04*</td>
</tr>
</tbody>
</table>

*** p-value < 0.001, ** p-value < 0.01, * p-value < 0.05

From the profile perspective, the user’s knowledge sharing behaviors were gender and social status (verified or not) independent. Individuals with less followers but more reciprocal relations were more likely to answer stranger’s questions routed to them. Figure 5-8 shows the differences in the number of followers and reciprocal relations for contributors and non-contributors in the data set.

![Figure 5-8](image)

Figure 5-8. Difference in log scaled number of followers and reciprocal relations for active contributors and non-active contributors.

I noticed that regarding the posting behaviors, non-contributors tended to retweet more than contributors, whereas, contributors interacted more with others by adopting a larger amount
of @mentions in their post. Compared with non-contributors, contributors were more willing to share cross-platform information with other, given that a high parentage of their tweets contained URLs. I demonstrate those differences in boxplots as shown in Figure 5-9.

![Figure 5-9. Difference in percentage of mention, URL sharing, and retweet posts for active contributors and non-active contributors.](image)

With the psycho-linguistic features, I observed that user’s contributing behaviors significantly associated with a set of LIWC measures. First, I found that posts of non-contributors contained more personal pronouns, such as I, we, you she, he, etc., indicating the more mundane nature of their posts. However, contributors used more impersonal pronouns, such as it, it’s, those, etc. As suggested by Ravid, van Hell, Rosado, and Zamora (2002) there was a positive correlation between the usage of impersonal pronouns and expository text. In addition, I also observed more verbs and fillers wrote in contributors’ Weibo posts. Prior studies on personality traits implied that high extraverts use more verbs to make their descriptions more lively (Oberlander & Gill, 2004). Third, contributors also used more words to describe their underlying cognitive mechanisms, including cause, insight, and tentative expressions. Fourth, contributors was
negatively associated with the use of work and achieve words, which according to Schwartz’s theory of human values (1992) reflect high self-enhancement.

Lastly, from the social activity perspective, I found that posts of contributors tended to receive more likes whereas less retweets than posts of the non-contributors. In other words, contributors were good at maintaining weak social relationships than non-contributors. Figure 5-10 depicts such differences.

![Graphs showing differences in percentage of posts received like, retweet, or any type of interaction for active contributors and non-active contributors.](image)

Figure 5-10. Difference in percentage of posts received like, retweet, or any type of interaction for active contributors and non-active contributors.

**Discussion and Implications**

In this work, I observed some interesting differences between Weiwen and traditional community Q&A services from three different perspectives, including user answering behaviors, interests, and social connectedness. First, I identified more contributors than consumers on Weiwen due to the question routing mechanism, which yielded a much higher response probability and response number than social Q&A in natural settings. Second, the participation in
Weiwen was unbalanced, where a small number of individuals contributed to a large proportion of questions. Third, the respondent’s topical interests increased along with their response frequencies, with more contributed individuals answered questions under more topical categories. Fourth, through a network analysis, I found that users in social Q&A sites based on question routing mechanisms seemed less connected than users in traditional Q&A settings. A more detailed analysis within each topical category further indicated the existence of connected communities under topics of Vexation, Life and Healthcare. I thought this might because of the common ground existed between individuals sharing either the same living background, or physical and emotional conditions under those topics.

With my findings on the well-separated roles of contributors and consumers on Weiwen, I further proposed a research question sought to explore the differences between contributors and non-contributors, assuming that they both had the expertise required to respond. To address this problem, I built a more classification model based on a set of non-QA features. Although overall the classifier demonstrated that using only non-QA features cannot very accurately predict a potential contributor in social Q&A, it still performed much better than a random guess (50%). Besides, since I tried to infer individual’s willingness to contribute from their intrinsic values (e.g. kindness and responsiveness), I thought the overall classification performance was comparable to other personality prediction studies. By analyzing each significant predictor in the regression model, I noticed that less popular but more interactive individuals on Weibo actually contributed more in social Q&A. Besides, users with more original and URL sharing posts answered more questions than those who retweeted a lot. From the psycho-linguistic perspective, I identified that individuals adopted less pronouns and achievement-related words, but more verbs and cognition-related expressions tend to contributed more when receiving a stranger’s question.

The contribution of this work is two-fold: first, I attempt to evaluate the effectiveness of the question routing scheme in social Q&A process. Although a number of previous studies
(Baichuan Li & King, 2010; Zhou et al., 2012) have suggested the adoption of question routing scheme in Q&A sites, such as Yahoo!Answers and even SNS, to the best of my knowledge, this is one of the first work that actually evaluate the performance of question routing services in real world settings. As my results suggested, there are both pros and cons to the use of question routing scheme in social Q&A, as it effectively increases the response rate of questions posted, however in the meanwhile decreases the social interactivity between users on SNS. Second, while identifying proper individuals for question routing, this study has a quite different perspective than most of the prior works (X. Liu et al., 2005; Pal, Chang, & Konstan, 2012). Instead of assessing individual's capabilities with respect to their answering probabilities, in this study I proposed a predicting model from the perspectives of individual’s desire in helping others when facing questions that fit their expertise. Just because I think that capability is not the only factor that determines one’s answering behavior in social Q&A.

In analyzing my results, I am aware of certain limitations that may restrict the ability to generalize the conclusions. One limitation is that this study is based on only one Chinese social Q&A service, Weiwen, so it may not be representative of the question answering behaviors demonstrated in other social Q&A sites, especially the cross-country services, such as Twitter and Facebook. However, J. Yang and Wei (2009) in their study analyzing the question and answering behaviors happened on Baidu Knows, Chinese largest Q&A sites, demonstrated similar characteristics as those reported for Yahoo!Answers (Adamic et al., 2008; J. Zhang et al., 2007). Therefore, I expect the same transition to be happened with Weiwen, although I would certainly like to apply my analysis methods on varied platforms. In addition, another limitation is that only non-QA features are adopted in this study to avoid the cold start problem. In future work, I would like to further improve the classifier by involving Q&A based features, such as previous answers provided, best answers provided, etc. In addition, I would like to research on incentive
mechanisms, which was currently adopted in Weiwen, to motivate non-contributors to share their knowledge with others.

**Conclusion**

To explore the knowledge contributing behaviors among strangers in social Q&A, in this study I experimented with Weiwen, a third party social Q&A application based on Sina Weibo. Different from traditional Q&A services, such as Yahoo!Answers and Baidu Knows, Weiwen was built with a question routing algorithm, which automatically directed received questions to individuals with required knowledge to respond. In order to understand the effectiveness of question routing in social Q&A environments, in this work I analyzed the Q&A activities on Weiwen collected during a ten months period. I noticed that question routing services such as Weiwen brought both pros and cons to social Q&A process, as it increases response probability, but at the meantime reduces the social interactivity between “friends” on SNS.

Besides, one of the problems with such question routing system was that, even if I know a person who had the expertise and experience to answer a question, their willingness to help was still indeterminate. To solve this problem, I also propose a predicative model automatically identifying active individuals who are willing to answer other’s questions. I built the model using non-QA features from four dimensions: profile, posting style, psycho-linguistic characteristics, and social activities. I adopted only non-QA features in the model considering the cold start situations. I found that from what an individual posts (e.g. mention, retweet, URL sharing, etc.) and how he/she posts it (e.g. usage of pronouns, achievement-related words, feeling-related words, etc), one could tell if the user is willing to answer a stranger’s question that fits his/her expertise.

I believe that this study provides a theoretical understanding of the knowledge contribution behaviors and patterns in social Q&A, as it provides both the advantages and
bottlenecks of the adoption of a question routing scheme in social Q&A. In addition, I view this work as a crucial step towards designing and implementing more accurate question routing mechanisms in social context, given its consideration on individual’s desires in answering questions, in addition to their capabilities.
Chapter 6
Conclusion

The objective of this thesis is to provide a conceptual understanding of social Q&A behavior and to provide insights into the design and development of technologies to support the information seeking process within social context. In order to achieve this goal, I have in this research proposed three studies, namely from the perspectives of user, information, and technology. Each of the aspects that I analyzed in this thesis corresponded to one dimension under the social Q&A framework as proposed in Shah et al. (2009)’s work. The results of these studies can lead to both theoretical and practical advances in information seeking under social context. This chapter summarizes findings in all three research studies, along with the contributions and ideas for future work.

Thesis Conclusion

Understanding user intent in Social Q&A

First, to understand the intentions behind questions posted on SNS, in Chapter 3, I presented the first study focusing on users involved in the social Q&A process. In this study, I proposed the following two research objectives:

Research Objective 1 (a): Develop a taxonomy of questions proposed in social Q&A that could be used to assist in selecting the most appropriate answering strategies.

To answer this question, I developed a taxonomy named ASK, which differentiate questions posted on SNS into three types, including: accuracy questions, in which people ask for
fact or common sense; social questions, in which people ask for coordination or companion; and knowledge questions, in which people seek for personal opinions or advices. In order to assess the validity of the taxonomy, I chose Twitter as the target platform, and recruited two annotators to work on the task of assigning 3000 randomly collected Twitter questions into the categories of accuracy, social, and knowledge. An inter-rater agreement of 87.37% indicated the relatively high reliability of the ASK taxonomy.

**Research Objective 1 (b): Implement the proposed taxonomy by automatically classifying questions into ASK types and measure the effectiveness of the classification.**

This question within the first study aims to explore the distinction between accuracy, social, and knowledge questions in the way they are being asked and answered. To measure the differences, I introduced features from five different aspects, including: lexical, topical, contextual, syntactic, and interact and social. I then trained and tested a prediction model using the 3000 annotated Twitter questions. Through feature analysis, I found that lexical features had the highest performance in terms of intention prediction. For instance, questions containing words such as “who”, “me”, and “want” were of high probability of being with social intent, whereas questions with “I” and “where” tended to be knowledge seeking. In addition, results showed that a majority of questions under the topical categories of “Beauty”, “Food”, and “Travel” are of knowledge-seeking intent. By incorporating features from the answer’s perspective, I further noticed that knowledge questions experienced longer time-lags in getting their initial answers and also tended to expire in longer durations. With all features taking into consideration, the predictive model achieved a classification accuracy of 83.20%.
Analyzing Question Response Probability in Social Q&A

Second, to solve the problem of no guaranteed response in social Q&A, in Chapter 4, I presented the second study focusing on factors influencing the question response probability. In this study, I proposed the following two research questions:

**Research Objective 2 (a):** *Identify the extrinsic factors that are likely to influence the question response rate in social Q&A.*

To answer this research question, I proposed in total 24 predictors that might affect the response probability in social Q&A. I tested the set of factors using a logistic regression model based on 62,106 questions collected from Weibo. Findings from this study indicated that factors, such as post originality, user popularity, social activities (like, retweet, comment), as well as question phrasing, all have positive effect on the response rate of questions asked in social context. However, the adoption of hashtag and first person pronouns both has negative influences on the chances of receiving answers.

**Research Objective 2 (b):** *Build a model to predict the question response probability using the extrinsic features.*

In addition to looking into each of the individual features, I also assessed the predictive power of the whole model in order to answer the second research question. I conducted a prediction experiment using the same features as proposed in the aforementioned logistic regression. The prediction model achieved a satisfactory accuracy of 0.74 in predicting question with or without responses. The analysis results indicated that whether or not a question would be
answered on Weibo was more related to the characteristics of the questioner, rather than the content of the question. In addition, to my surprise, among all the user-based features, social activities showed much higher discriminative power in predicting response probabilities under social context.

**Characterizing Knowledge Contribution in Social Q&A**

Third, to assess the effectiveness of existing social Q&A technologies, in Chapter 5, I presented the third study characterizing the knowledge contribution behavior within a question routing service called Weiwen. In this study, I proposed the following two research questions:

**Research Objective 3 (a):** Explore the question answering patterns of individuals when they are exposed to questions asked by strangers via question routing.

To explore the patterns of knowledge contribution within question routing frameworks, I examined the differences between Weiwen and traditional community Q&A services from three distinct perspectives, including user behaviors, interests, and connectedness. I noticed that question routing services, like Weiwen, were effective in attracting responses from a wider range of audience. To be more specific, though the analysis I noticed more contributors than consumers on Weiwen due to the question routing mechanism, which yielded a much higher response probability and response number than social Q&A in natural settings. Second, the participation in Weiwen was unbalanced, where a small number of individuals contributed to a large proportion of questions. Third, the respondent’s topical interests increased along with their response frequencies, with more contributed individuals answered questions under more topical categories. However, I also observed some limitations that existed in the current design of question routing
services. Through a network analysis, I found that users in social Q&A sites based on question routing mechanisms seemed less connected than users in traditional Q&A settings, with a well-separated roles of knowledge contributors and consumers.

**Research Objective 3 (b):** Identify individuals with the desire to help others in social Q&A by using their non-Q&A characteristics.

To explore the distinction between active users and non-active contributors in social Q&A, I proposed a predictive model based on non-QA features from four dimensions: profile, posting style, psycho-linguistic characteristics, and social activities. I trained and tested the classification model based on 3,006 active and non-active contributors extracted from the collected data set. Overall the classifier demonstrated that using only non-QA features can very accurately predict a potential contributor in social Q&A. By analyzing each significant predictor in the predictive model, I observed that less popular but more interactive individuals on Weibo contributed more in social Q&A. Users with more original and URL sharing posts answered more questions than those who retweeted a lot. From the psycho-linguistic perspective, I identified that individuals adopted less pronouns and achievement-related words, but those posters using more verbs and cognition-related expressions tend to contributed more when receiving a stranger’s question.

**Research Contributions**

First, from a theoretical point of view, given the limited number of studies on social Q&A, this research can be regarded as a pioneer attempt to comprehensively explore patterns demonstrated in user participations in order to understand social Q&A behavior. The results from
this research can be of great help in understanding the issues exist in the current design of SNS and question routing services in handling question-answering. In addition, results from this thesis can also contribute to the body of current knowledge regarding social Q&A and can be adopted as the foundation for future studies with the related focus.

Second, this research is also of practical values in designing and developing systems or tools to facilitate the social Q&A process. To be more specific, with the proposed ASK taxonomy, along with the predictive model on user intent, systems could be designed to automatically categorize posted questions into accuracy, social, and knowledge types. According to the different underlying intents behind the three types of questions, systems could then rely on the archived knowledge to automatically answer the accuracy-seeking questions, and route those knowledge-seeking questions to audience with similar background or experiences. In addition, with the predictive model on response rate, as presented in my second study, social Q&A systems can check to see in advance if a question has the potential to receive a response. If so, the system can just wait for someone from the questioner’s network to answer that question, if not, then the system can direct that question to a larger audience to save time and effort. Finally, my third study allows the question routing system to select potential respondents not only based on expertise but also based on their internal traits, such as willingness and kindness to help.

**Future Research**

I am interested in continuing this line of research on modeling user behavior in social Q&A, as well as developing tools to support the social information seeking process. In this section, I list a number of potential directions for future research within this area.
Running Experiment based on the Current Findings

Since all of the exiting works in my thesis were based on data analysis, although using real world data enabled us to gain a conceptual understanding of the social Q&A activities, still I want to test my findings via user studies. Because according to Jansen (2009), trace data will not interfere with the natural flow of behavior in real world contexts. Possible directions for user studies include: 1) Assessing the predictive model of active contributors in real-world setting. In order to do that, I will first identify potential contributors based on the model as presented in my third study. After manually send questions to those predicted contributors, I could then evaluate the effectiveness of the predictive model by comparing the perceived responses with the predicted results. 2) Identifying intrinsic factors that influence individual’s knowledge contribution in social Q&A. I will survey the potential respondents for their internal traits, including their big 5 personality and their values. Then based on the survey results, trying to see if one’s internal traits could affect their willingness to offer help in social context.

Improvement on the Question Routing Models

In addition to optimizing and expanding my current works, I also plan to conduct further studies to improve the design of the question routing models. Question routing has been studied intensively on the platform of community Q&A sites (Baichuan Li & King, 2010; Pomerantz et al., 2003; Zhou et al., 2012). However, given the dissimilar experience of asking questions on community Q&A sites versus on SNS (Morris et al., 2010b). I believe that question routing on those two platforms could also be different, and thus needs to be studied separately. In addition, I noticed that most of the question routing methods existed were expertise-based. However, as I indicated in the second study, different from the community Q&A services, responses in social
Q&A were more “friendship-driven” or “social-driven”, rather than “expertise-driven”. So, in that sense, more features besides individual’s expertise need to be considered when designing a question routing mechanism for social Q&A sites. Some of the potential factors that could be considered in designing a question routing model include: 1) Location, as I indicated in this thesis, spatial context seemed extremely important in social context than in community Q&A sites. 2) Availability, in addition to the spatial context, temporal constraints of a question is also important in social Q&A. In addition to developing a way to quantify the urgency of the questions, predicting the availability of a potential respondent could be equally important in designing effective question routing mechanisms. 3) Social Tie Strength, given the social nature of SNS, social tie strength should also be added into the model of question routing. In my third study, I noticed that the current design of question routing services for Weibo generated relatively disconnected communities, which deviated from the nature of social Q&A. Because according to Morris et al. (2010b), people prefer strong social ties for question answering in social context, and trust them more than those weak ties.

Development of a Mobile-based Social Q&A Platform

In addition to an improved model of question routing, another interesting direction for social Q&A could be migrating the current systems and tools to a mobile platform. To my understanding, a mobile environment could benefit the social Q&A activities in a number of ways. First, according to the recent announcement, 80% of Weibo active users are on mobile platforms (China Internet Watch, 2014). By targeting on those highly active users, questions asked mobile platforms could have higher chances of being answered than questions asked on SNS. Second, with the help of various sensors implemented in portable devices, mobile applications can collect valuable information (e.g. location, availability) that is difficult to obtain via web browsers. With
this user-related information one can build more accurate and efficient models for social information seeking.

Closing Remarks

Social Q&A, as a newly emerging environment for information seeking, has been lately investigated by a number of studies. As a comprehensive analysis of information seeking under social context, this thesis explored social Q&A from three aspects: user, information, and technology. Through data analysis, I examined 1) the types of questions asked in social Q&A; 2) the extrinsic factors that influence the social Q&A response; and 3) the effectiveness of question routing services within social Q&A frameworks. I also proposed models that could be used to 1) categorize questions according to their underlying intents; 2) to estimate the response probability of a question; 3) to identify the active contributor based on their SNS feeds. The result of this thesis could be of specific interest in the question and answering area; in particular those who want to design and develop systems or tools to facilitate the social Q&A processes.
References


Dror, G., Koren, Y., Maarek, Y., & Szpektor, I. (2011). I want to answer; who has a question?: Yahoo! answers recommender system *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1109-1117). San Diego, CA, USA: ACM.


Curriculum Vitae

Zhe Liu

EDUCATION

Ph.D. in Information Sciences and Technology  
The Pennsylvania State University  
Advisor: Bernard J. (Jim) Jansen  
May, 2015

Master in Management Information Systems  
University of Arizona  
May, 2009

Bachelor in Management Information Systems  
Beijing Language and Culture University  
June, 2007

PEER-REVIEWED PUBLICATIONS

Journal Papers


Conference Full and Short Papers


