8-2012

SOCIALQ&A: A NOVEL APPROACH TO NOTIFIYING THE CORRECT USERS IN QUESTION AND ANSWERING SYSTEMS

Nikhil Vithlani
Clemson University, nvithla@g.clemson.edu

SOCIALQ&A: A NOVEL APPROACH TO NOTIFIYING THE CORRECT USERS IN QUESTION AND ANSWERING SYSTEMS

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Computer Engineering

---

by
Nikhil Vithlani
August 2012

Accepted by:
Dr. Haiying (Helen) Shen, Committee Chair
Dr. Melissa C. Smith
Dr. Jill Gemmill

ABSTRACT

Question and Answering (Q&A) systems are currently in use by a large number of Internet users. Q&A systems play a vital role in our daily life as an important platform for information and knowledge sharing. Hence, much research has been devoted to improving the performance of Q&A systems, with a focus on improving the quality of answers provided by users, reducing the wait time for users who ask questions, using a knowledge base to provide answers via text mining, and directing questions to appropriate users. Due to the growing popularity of Q&A systems, the number of questions in the system can become very large; thus, it is unlikely for an answer provider to simply stumble upon a question that he/she can answer properly. The primary objective of this research is to improve the quality of answers and to decrease wait times by forwarding questions to users who exhibit an interest or expertise in the area to which the question belongs. To that end, this research studies how to leverage social networks to enhance the performance of Q&A systems. We have proposed SocialQ&A, a social network based Q&A system that identifies and notifies the users who are most likely to answer a question. SocialQ&A incorporates three major components: *User Interest Analyzer*, *Question Categorizer,* and *Question-User Mapper*. The *User Interest Analyzer* associates each user with a vector of interest categories. The *Question Categorizer* algorithm associates a vector of interest categories to each question. Then, based on user interest and user social connectedness, the *Question-User Mapper* identifies a list of potential answer providers for each question. We have also implemented a real-world prototype for SocialQ&A and analyzed the data from questions/answers obtained from the prototype. Results suggest that social networks can be leveraged to improve the quality of answers and reduce the wait time for answers. Thus, this research provides a promising direction to improve the performance of Q&A systems.

DEDICATION

I would like to dedicate this thesis to my mother Hansa Vithlani, who was an example that with hard work and dedication anything is achievable and without whom even school education would not be possible for me; to my father Jayantilal Vithlani, who always believed in me and imbued me with will power; to my grandfather Jaman Vithlani, without whom coming to United States of America and doing my Master's would still be a dream; my brother Riten Vithlani, who always encouraged and motivated me; and all my wonderful friends who believed in me more than I did in myself, which inspired me to do more than I thought I could. Finally, I would like to dedicate this thesis to Dr. Haiying Shen, who helped me all throughout my Master's and gave me the opportunity for research. I am very thankful for her time, support, help, co-operation and lot more. I can't thank her enough.

ACKNOWLEDGMENTS

While writing this thesis I had the support and encouragement of all my professors, colleagues, friends and family. I would like to extend sincere thanks to all of them.

I would like to thank my advisor and committee chair Dr. Haiying Shen for her support and insights all throughout my Master’s work. I would like to thank my committee members Dr. Melissa Smith and Dr. Jill Gemmill for their valuable advice and time for helping me significantly improve my thesis. I am very grateful for their help. I am thankful to Dr. Poole for his guidance, motivation and directions. He helped me keep my spirits high.

I would like to thank my colleagues in my research group, Ze Li and Guoxin Liu, for their encouragement and advice.

I would like to thank my friends, Michael Green, Timothy Reeves, Brian Von Fange and Harrison Chandler for their insightful editorial advice. Also, I would like to thank Josh Mitchell and Michael Green for their support and motivation all throughout my study in the US, they have helped me as my brothers and I really feel indebted to them.

TABLE OF CONTENTS

Page

CHAPTER

Table of Contents (Continued)

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

List of Figures (Continued)

Figure Page

CHAPTER 1

INTRODUCTION AND MOTIVATION

In this chapter, we motivate and introduce the research. We present the background for Question and Answering (Q&A) systems, the motivation for developing a new Q&A system, and the objectives and contributions of our research.

The Internet is an important source of information, and the amount of data on the Internet is vast and constantly growing. Users rely on search engines to find specific information within this knowledge base. Search engines such as *Google*[1] and *Bing*[2] do a good job of indexing web pages and providing users with pages relevant to their search queries. These search engines use keywords provided by the users to perform searches; however, there are some specific questions that are not suited for search engines. For example, "Where is the best place to get your car fixed in Clemson?" Q&A systems have been developed to address this particular class of non-factual questions. Since their inception, Q&A systems have proved to be a valuable resource for sharing expertise and consequently are used by a large number of Internet users.

[1] http://www.google.com

[2] http://www.bing.com

Q&A systems also preserve all previous questions and answers, thus acting as a repository for information retrieval. Currently, Q&A systems play a vital role in academia, as they can aid students who use online learning systems to resolve their questions. Many students post their questions on online Q&A systems such as *Yahoo! Answers*[3] and *Stackoverflow*[4]. As mentioned by Adamic *et al.* [55], Q&A sites are not only important for sharing technical knowledge, but also as a source for receiving advice and satisfying one's curiosity about a wide variety of subjects. Due to the growing importance of Q&A systems, many researchers have focused on improving the functionality and efficiency of Q&A systems. As mentioned by Radford *et al.* [3], the growing importance of Q&A systems in both research and academic communities demands an effort to better understand these systems and strive towards improving them. Hence, it is important to contribute to the improvement of Q&A systems.

In this thesis, the term "end user" represents a user who posts a question, the term "answer provider" represents a user who is considered to have the potential to provide an answer, and the term "user" represents any general user in the system.

There are many Q&A systems available such as *Yahoo! Answers*, *StackExchange*[5], *Quora*[6], etc. These are widely used by vast populations on

[3] http://www.answers.yahoo.com
[4] http://www.stackoverflow.com
[5] http://www.stackexchange.com
[6] http://www.quora.com

a daily basis: Yahoo! Answers was launched at the end of the year 2005 and had more than 10 million users as of February of 2007 [4], and according to the Yahoo! Answers blog, there are currently 200 million users with 15 million visits every day [5]. This shows that the number of users is increasing exponentially and also that these users are active.

Current Q&A systems consist of hundreds of thousands of users, so the number of questions asked is also very large. Consequently, when a user intends to answer a question, he/she may be overwhelmed by the plethora of questions needing answers. Moreover, there are potentially some questions where a user has expertise and can provide a better answer than other users, but there is currently no way for him/her to locate those particular questions among the thousands of posted questions. For a given question, the user who is interested or has expertise in a specific topic would provide better answers than the user who possesses less knowledge of the topic. Thus, there is a need to develop a mechanism that would forward questions to the appropriate answer providers, whose interest/expertise matches the question's topic(s).

To map questions to answer providers, currently available Q&A systems allow end users to choose tags (interest categories) for their questions. However, such an approach has two problems:

1. The tag(s) provided by the end user might be inaccurate.
2. Sometimes, the end user does not know the appropriate tag(s) should to attach to a question.

Li *et al.* [5] carried out research on routing questions to the appropriate users; they tracked 3000 random questions from Yahoo! Answers and Baidu Zhidao for a period of 48 hrs. As shown in Figure 1, they found that for Yahoo! Answers, only 17.6% of questions were answered satisfactorily. From the remaining 82.4%, one fifth of the questions remained unanswered. For Baidu Zhidao, 22.7% of questions were successfully answered, and 42.8% of the unresolved questions were not answered at all [5]. Clearly, there is room for improvement in the Q&A domain to decrease the number of unanswered questions in a Q&A system. Hence, there is an increasing need for an advanced method to route questions to those users with the highest likelihood of answering them with expertise in that subject area.

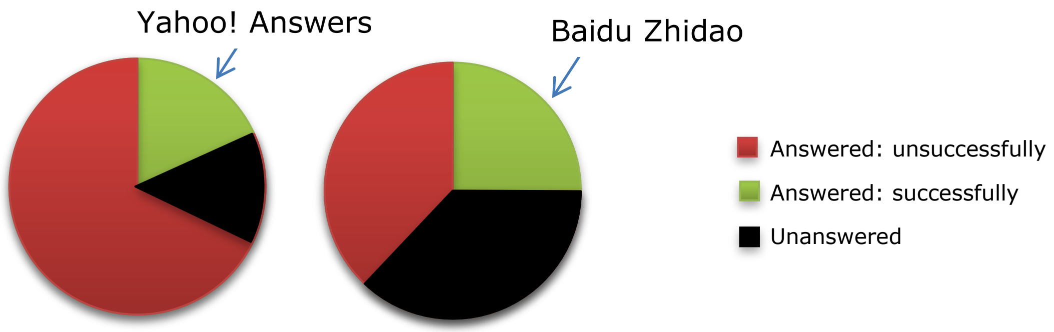


Figure 1. Q&A statistics related to Yahoo! Answers and Baidu Zhidao based on data provided by Li *et al.* [5].

Towards this goal, this research studies leveraging social networks to route questions to appropriate answer providers to improve the quality of answers provided by the answer providers and to reduce the amount of time the end user must wait to obtain an answer. We propose a Q/A system called SocialQ&A that considers user interest and social connectedness to identify potential answer providers that would provide high-quality answers in a short time period. Though previous research efforts [2,64] also use social networks for Q&A systems or search engines, this research is different from previous efforts in two aspects: (1) it aims to improve the quality of answers and reduce the wait time for answers, and (2) it explores a different method to identify potential answer providers for questions. SocialQ&A derives each user's interests from his/her profiles and Q&A activities, and produces the user's interest vector. It also calculates the social connectedness between users based on their interest similarity, interactions and common friends. To identify potential answer providers, SocialQ&A considers two metrics: the interest of the answer provider towards the question and the social connectedness of the answer provider with respect to the end user.

The contributions of this Master's thesis are as follows:

1) The design of SocialQ&A. SocialQ&A is a social network based Q&A system developed as a part of this research. SocialQ&A is composed of three components: 1) *User Interest Analyzer*, 2) *Question Categorization,* and 3) *Question-User Mapper*.
2) The implementation of a real-world SocialQ&A system. We have prototyped the SocialQ&A system and conducted a real-world test

with 124 users from India, the United Kingdom, and the United States for a period of approximately one month.

3) The collection and analysis of the data from SocialQ&A. We have analyzed the features of the questions posted, the questioning and answering activities of users, the quality of questions, the wait time for answers, and the question categories.

It is indicated in [63] that Computer Engineering is the design and prototyping of computing devices and systems, and concentrates its effort on the ways in which computing ideas are mapped into working physical systems. One main branch of Computer Engineering is "Networks" that is concerned with design and implementation of distributed computing environments, from local area networks to the World Wide Web. This research focuses on the design and prototyping of a working physical system, a social network based Q/A Q&A system.

In this chapter, we have introduced Q&A systems and the motivations for this research. We briefly described our proposed system and highlighted our contributions. The remainder of this thesis is organized as follows. Chapter 2 covers the background of Q&A systems, the history of search engines, information retrieval paradigms, and the evolution of Q&A systems over time. It also provides a brief overview of related research conducted in Q&A systems. Chapter 3 explains the architecture and implementation details of SocialQ&A. Chapter 4 provides the testing results and analysis obtained from the real-world SocialQ&A prototype. Chapter 5 offers conclusions and potential future research directions.

CHAPTER 2

BACKGROUND AND RELATED WORK

The purpose of this chapter is to familiarize the readers with the background of Q&A systems and state-of-the-art studies on SocialQ&A systems. Q&A systems are closely related to search engines and information retrieval paradigms. Thus, we first introduce the history of search engines and information retrieval paradigms, and then introduce the evolution of Q&A systems concerning the shift towards social searches, text mining approaches, and answer provider identification. As answer provider identification is the most relevant topic to our research, we present a review of previous studies on this topic and briefly present the distinguishing features of our proposed SocialQ&A system.

2.1 Background

In this section, the motivation for development of Q&A systems and the evolution of Q&A systems are discussed in detail. The section also describes two commonly implemented information retrieval paradigms, namely, the Library paradigm and the Village paradigm [2]. Finally, it provides the preliminary concepts that were used to implement SocialQ&A.

2.1.1 The history of search engines

There is an abundance of web sites present on the Internet with vast amounts of information, each of which is increasing rapidly. With so much information present on the Internet, it can be problematic for users to find specific information. This information retrieval problem was the basic incentive for the invention and development of search engines [1,32]. Initially, search engine databases were constructed manually; thus, they were difficult to maintain and update. As mentioned by Brin and Page [1], these search engines sufficiently indexed the most interesting and common topics, but failed to collect information that was uncommon and sparse.

Until the arrival of Google, automatically indexed search engines were considered substandard because of the low quality of search results that they returned. Google, which originated from the Stanford Digital Library Project by Page *et al.* [22], transformed the way automated search engines worked by making the search process extremely intelligent, thus eliminating the noise in the search results that had been present in earlier automated search engines. Google makes heavy use of the additional structure present in hypertext to provide higher quality search results and is designed to scale well for extremely large data sets. It also makes efficient use of storage space to store the index, and its data structures are optimized for fast access [1]. However, with the passage of time, searching using only web-based search engines for a specific query became a tedious task because the queries of users were in natural language, but the search engine tried to use keywords from the query to find a relevant web page, assuming it would provide the user with the desired answer. Since traditional search engines perform poorly when the question is asked in natural language, the challenge of natural language queries laid the foundation for online Q&A systems and an entirely new area of research in the field of online computing.

2.1.2 Information retrieval paradigms

The most fundamental and widely adopted paradigm for information retrieval is the Library paradigm as described in [2], which is used by Google and most other contemporary search engines. The Library paradigm uses keywords as the criteria for searching. The information is present in the form of web pages and the user provides various keywords relevant to his/her query to a search engine. The search engine, in turn, provides related web pages to the user. The web pages are indexed by an administration authority such as Google, Yahoo!, Microsoft, etc.; thus, the trust is based on authority. The algorithm implementing the Library paradigm is designed to use the cues provided by the end user in the form of search keywords to calculate the relevance of a web page to those words. The relevant web pages are then represented to the user as search results. It is the task of the user then to find the correct web page from these results.

The name 'Village' in the Village Paradigm [2] comes from the way information retrieval functioned before the Internet era. In a village, people used natural language to ask questions and directed those questions to people who they knew personally and who would be able to answer the questions. Thus, the flow of information in the Village paradigm is based on the social connections of the user.

The objectives of systems employing the Library paradigm are very different from systems employing the Village paradigm. The main aim of a system using the Library paradigm is to find the web page that can provide the appropriate information pertaining to the search keywords specified by the user, whereas the goal of a Village paradigm system is to find the appropriate person rather than the appropriate web page [2].

Neither paradigm is perfect; however, each paradigm has certain scenarios that make one more useful than the other. For example, if an individual wants to know the area of a country or the population of a country, the Library paradigm would be more suitable since one would not expect his/her friends or colleagues to remember such facts. Conversely, if one would like to know a good course to take during the spring semester at a given university, the Village paradigm would be more useful than the Library paradigm. Thus, certain questions are inherently ill-suited for the keyword search approach, because people tend to consult others in matters of opinion. However, the strength of the Library paradigm for information retrieval relative to the Village paradigm is that the end user does not depend on another individual for the resolution of his/her query. In summary, the Library paradigm is more suitable for fact-based questions, and the Village paradigm is more suitable for opinion-based questions.

2.1.3 Evolution of Q&A systems

Q&A systems provide a web-based environment for users to communicate with each other. The end users ask questions in natural language through the user interface. The question is visible to all users in the system, and the users that have enough expertise to answer a question do so. There can be multiple answers to a single question; subsequently, the end user can decide which answer is the best for his/her question. Q&A systems are also a useful resource for the reuse of the acquired information, since the questions that are answered successfully are stored in the system and other end users with a similar question can search the database to obtain the solution immediately.

2.1.3.1 Shift towards social searching

The Village paradigm has resulted in an evolution of Q&A systems. Evans *et al.* [37] identified searching as a social activity, as opposed to a solitary activity, and demonstrated that social interactions before, during, and after the search activity can help improve the search results. Morris *et al.* [38] discussed the growing trend towards posting queries as social network statuses instead of using web search engines.

Moreover, as stated by Barker [17], there has been a shift in the world of education in last two decades towards making the process of learning based on constructivism rather than on transmission, that is, toward making the learning environment move from teacher-centered to student-centered.

According to Putnam *et al.* [18], the educators and teachers are now basing their teaching methodologies on principles of social learning where learning takes place in collaboration instead of in isolation [19]. In the online learning environment, students ask questions regarding the material they have learned and experts in a particular domain provide a very useful resource for aiding these students, as stated by Han *et al.* [16]. Thus, Q&A systems could be a very useful accessory to online learning environments.

2.1.3.2 Text mining approaches

Since the advent of Q&A systems, there have been many attempts to improve the quality of the answers provided to the questions and to minimize the time period involved between the posting of a question and the response to the question [2,5,6,7,8,10]. There has been much research focused on making Q&A systems intelligent such that they can provide answers automatically without the need of human users [56,39,41,40]. In these systems, previous answers given by human users are used as the knowledge base to form the new answers. The research by Akiyoshi *et al.* [11] was based on the algorithms used for retrieval of similar Q&A articles in web bulletin boards. The authors believe that the methods presented by Mochihashi *et al.* [24], Radev [25], and Sakurai *et al.* [26] are similarity-based methods and do not utilize the best information present in the end user's query. They proposed a method where they obtained the relevance index from commercial web search engines. Relevance index is a measure of how closely the thread in the web bulletin board is related to the end user's search query. To calculate the relevance index using Internet search engines, they take a ratio of AND retrieval and OR retrieval for all the words present in the search query and use this ratio to determine the association index. This research improves the retrieval accuracy over that of keyword-based retrieval, but is inefficient due to the unrelated and irrelevant keywords present in the articles. The method presented by Akiyoshi *et al.* [11] exploits the inherent structure of bulletin board systems (BBS), which have a thread structure containing one query and multiple solutions to that query. The algorithm compares the association index from web search engines to the relevance index derived from the BBS structure. Based on the experiments conducted, the algorithm in [11] improves the retrieval accuracy by 30% compared to a similarity-based method.

Research by Xie *et al.* [21] is directed toward mining information from web pages and presenting the mined answers to the end user. The authors claim that prior work in this area focused on returning answers related to the question asked by the end user. However, the authors identified that those answers were not accurate, therefore, they were not as useful to the end users. The authors proposed an alternative solution where the objective is to perform the semantic analysis of all answers returned by the search engines and then present the end user with a fused answer. The objective is to fuse the answers based on their similarity. The answers are clustered using a lexical database like *WordNet*[7] and then the answers belonging to the same cluster are fused. The fusion is carried out using three methods: 1. Data quality-based fusion, which uses WordNet to determine the quality and then assigns data quality attributes to an answer dynamically; 2. Content rule-based fusion, where the users rate the answer using 11 predefined tags like min, max and major; and 3. Mixed method-based fusion, which considers both the first and the second methods (details and mathematical representation can be found in [21]).

2.1.3.3 Answer provider identification

The authors of [57,58,59,60,64] concentrate on locating experts and authoritative users in the system. Much research in Q&A systems was directed toward the categorization of questions into pre-defined categories [42, 43, 44], making it easier for end users to locate previously asked questions as well as for experts to find questions they can answer.

Some systems use a reputation system to depict the credibility of the answer provided by the user. Users providing high-quality answers would be rated higher by his/her peers and thus, would have good reputations. On the other hand, users providing answers that are not at all useful or are of mediocre quality would have relatively lower reputations. Consequently, studies are conducted to create a reputation model and incorporate that into

[7] WordNet is a registered trademark of Princeton University, available by anonymous ftp from clarity.princeton.edu.

Q&A systems [45,46] and to determine the relationship between the reputation of the user and the quality of answers provided [47].

2.1.4 WordNet

WordNet is a lexical database for the English language that is used for natural language system development [13,20]. English nouns, verbs, adjectives, and adverbs are organized into sets of synonyms, each representing a lexicalized concept; semantic relations link the synonym sets [14]. In WordNet, the words are not represented by their individual forms, but by their meanings or lexemes. The meaning of a word is described using a set of synonyms (Synset) that represent that word [15]. The path-lengths between words indicate synonymous proximity between the words. We use WordNet to parse the user information in SocialQ&A to derive users' interests. SocialQ&A uses one-hop path length for generating the *Synset* of each of the pre-defined interest categories. Figure 2 shows the two-hop Synset for the word "be" as generated by wordnet.

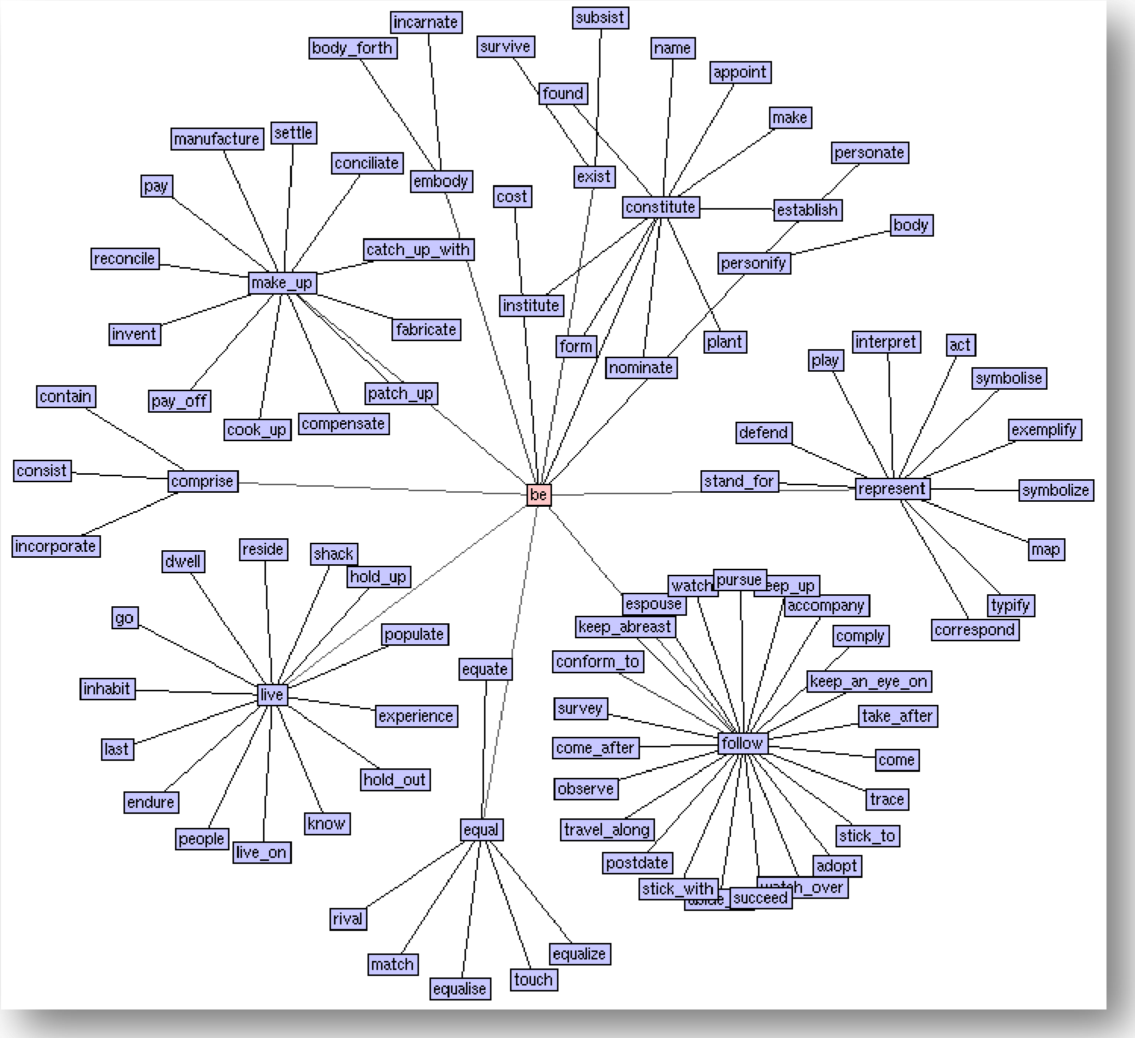


Figure 2. Two-hop Synset for the word “be” generated by WordNet.

2.2 Related work

Since identifying answer providers in Q&A systems is the most relevant topic to this research, we present a review of previous studies on this topic in this section.

Li *et al.* [6] conducted a study aimed at incorporating the concept of *question category* to question routing systems for improving the efficiency of community-based Q&A systems. The study focused on 400,000 resolved questions belonging to the 'Computer & Internet' and 'Entertainment & Music categories of Yahoo! Answers. They showed that including the concept of *question category* for question routing in community based Q&A systems can provide an answer provider expertise with higher accuracy compared to the traditional Query Likelihood Language Model (QLLM) proposed by Liu *et al.* [27], the state-of-the-art Cluster-Based Language Model by Zhou *et al.* [28], and a mixture of Latent Dirichlet Allocation and QLLM presented by Liu *et al.* [10]. Moreover, they showed that from a computing cost perspective, the proposed category sensitive language model is more efficient than the three models stated above. The paper presents detailed information regarding the degree to which the proposed method is superior to the other mentioned methods. Table 1 below contains the precision of the method proposed by the authors versus the other methods.

Table 1. Different methods' precisions in question routing, B. Li *et al.* [6].

K	QLLM	BCS-LM	TCS-LM	LDALM	CBLM
1	0.0795	0.1114 (↑40.13%)	0.1227 (↑54.34%)	0.0989 (↑24.40%)	0.0000
3	0.1659	0.2364 (↑42.50%)	0.2340 (↑41.05%)	0.1950 (↑17.54%)	0.0000
5	0.2091	0.2727 (↑30.42%)	0.2705 (↑29.36%)	0.2455 (↑17.41%)	0.0000
10	0.2705	0.3386 (↑25.18%)	0.3455 (↑27.73%)	0.3102 (↑14.68%)	0.0000
20	0.3386	0.3909 (↑15.45%)	0.3932 (↑16.13%)	0.3710 (↑9.57%)	0.0091
40	0.4136	0.4523 (↑9.36%)	0.4591 (↑11.00%)	0.4392 (↑6.19%)	0.0273
60	0.4477	0.4818 (↑7.62%)	0.4795 (↑7.10%)	0.4649 (↑3.84%)	0.0545
80	0.4727	0.4955 (↑4.82%)	0.4909 (↑3.85%)	0.4867 (↑2.96%)	0.0727
100	0.4909	0.5159 (↑5.09%)	0.5114 (↑4.18%)	0.4979 (↑1.43%)	0.0795

A general phenomenon seen in Q&A systems occurs when two or more end users ask the same question repeatedly; this condition is undesirable because it wastes system resources due to the presence of redundant information. Moreover, this is an annoyance for Q&A system users, since they see the same question asked repeatedly even when it has been answered in the past. Cao *et al.* [7] focused their research on improving the user's experience by decreasing the user wait time between asking a question and receiving an acceptable answer. The authors devised an algorithm that determines if a similar question exists among any previously asked questions when a user posts a new question. If the algorithm is able to determine with sufficient confidence that a similar question does exist, it suggests those questions and answers so that the end user does not need to wait and can benefit from the previous expertise. The authors exploited the category classification from the "Question Answer" archives of various communities and implemented a local smoothing algorithm to make their searching more efficient and accurate. Similar studies were conducted by Duan *et al.* [29], Jeon *et al.* [30], and Wang *et al.* [31] where the main focus of the research is to find similar questions in a community Q&A system. Additionally, there were attempts in the past to study the quality of answers provided in Q&A scenarios [52,53,54,58].

Horowitz *et al.* [2] proposed to make search engines social. They developed a social search engine known as *Aardvark* for this purpose. *Aardvark* is formed from four main components as discussed in the paper: (1) Crawler and Indexer, (2) Query Analyzer, (3) Ranking Function and (4) User Interface. The users of *Aardvark* can enter their search queries through a text message, an email, or a normal web browser. The queries are presented to *Aardvark* in natural language. The aim of this research is to make the process of searching more social by providing the users with a real-time system to communicate with one another mediated by *Aardvark*. *Aardvark*'s goal is to find a user who could potentially resolve the search query of the end user in real time. After finding the appropriate user, *Aardvark* determines whether this user could assist the end user, waits a pre-determined time for a response, and then moves on to the next appropriate user in the list until the end user receives a response. A total of 90,361 users tested *Aardvark* actively over the period of 6 months, and from those users, 78,343 provided feedback for the research where *Aardvark* was compared with Google search. It was found that 71.5% of the total queries were answered successfully on *Aardvark* with a mean rating of 3.93 out of 5; while 70.5% of the queries were answered successfully on Google with a mean rating of 3.07 out of 5. Thus, the research indicates that the average user satisfaction was higher for *Aardvark*, reflecting that the users found the quality and relevance of the answers in *Aardvark* to be better than the search results given by Google.

To efficiently identify potential answer providers, Li *et al.* [64] proposed a distributed Social-based mObile Q&A System (SOS) with low node overhead and system cost as well as quick response to questions. SOS leverages the lightweight knowledge engineering techniques to transform users' social information and closeness, as well as questions to IDs, respectively, so that a node can locally and accurately identify its friends capable of answering a given question by mapping the question's ID with the social IDs. The node then forwards the question to the identified friends in a decentralized manner. After receiving a question, the users can decide to forward the question or answer the questions if able. The question is forwarded along friend social links for a number of hops, and then resorts to the server. The cornerstone of SOS is that a person usually issues a question that is closely related to his/her social life.

Guo *et al.* [8] explored the topic of recommending potential answer providers. Their approach is to delineate a ranked list of potential answer providers by solving three associated sub-problems associated with this task. First, to tackle the problem of finding the focus of the question, they used two forms of question representation: topic-level representation and term-level representation. Topic-level representation is basically the same as categorization; on the other hand, for the purpose of term-level representation, the authors use the BM25F method [9], which is an extension of the 2-Poisson model of term frequencies in documents [61], described in detail by Robertson *et al.* [9]. The main advantage of using BM25F is that it preserves the term frequency information of the text. The second sub-problem described by the researchers is that of defining user expertise and interest representation. For this purpose, they used the topic-level description of a question and also the profile information of the users. Then, they mined for terms in questions as well as previous answers provided by the user in question to define his/her expertise and interest. The third sub-problem is that of ranking the potential candidates, which is tackled by assigning weights to topic-level similarity rank and the term-level similarity rank and combine them.

A similar study was done by Li *et al.* [5] two years later. They proposed a 'Question routing framework' wherein they disintegrated the process of routing a question into four phases: (1) Performance Profiling, (2) Expertise Estimation, (3) Availability Estimation, and (4) Answerer Ranking. When a question is posted, users are profiled based on their past answering performance. The next step estimates the user's expertise based on his/her interest and profile. For expertise estimation, a number of features are extracted from the answer, and Kernel Density Estimation [50] is used for conversion of non-monotonic features. Then, the potential answer providers are checked for availability based on their past login times. After taking these factors into consideration, the questions are routed to the most highly ranked users.

Liu *et al.* [10] focused on a similar problem but used a different approach to model users' interests. They also took user authority and activity into account for better results. The methods used in this research are: (1) Language Model and (2) Topic Model, based on Latent Dirichlet Allocation by Blei *et al.* [48]. The Language Model uses words appearing in the question and the words occurring in all previous answers to calculate the interest of the potential answer provider to answer the question, and then uses the Dirichlet smoothing method originally proposed by Zhai *et al.* [49]. The Topic Model based on Latent Dirichlet Allocation (LDA) approaches the problem of lexical gap, which is the weakness of the Language Model. The lexical gap problem is addressed by identifying the latent topic of interest for the potential answer provider. Using LDA, the words in the user profile are used to generate a corpus of words, which defines the user profile including the possibly latent interests of the user. This corpus of words is then used to estimate the probability that the user can/will answer the question.

2.3 SocialQ&A

The important difference between SocialQ&A and previous social network based Q&A approaches covered in the related work is that SocialQ&A

uses a different method to exploit the answer provider’s profile information and interests as well as the end user’s social network to route the question. Additionally, interest information for all users in the system is continuously updated based on their actions (questions they ask and questions they answer). SocialQ&A aims to improve the answer quality and reduce the wait time for answers. Unlike many prevalent Q&A systems, SocialQ&A routes the questions only to the answer providers in the end user’s social network to ensure that the notifications do not become a source of frustration for answer providers. However, any user can still see all questions asked by any end user of SocialQ&A by browsing the recently posted questions, regardless of how the questions were routed. Any user can also answer or forward a question regardless of whether it was specifically routed to him/her by the system.

2.4 Summary

In this chapter, we have discussed the development and evolution of Q&A systems in the quest to provide answers to questions asked using natural language by end users. There has been continuous innovation in the field of Q&A systems, and we have reviewed many studies conducted for the improvement and development of Q&A systems. We have identified the unique contribution of SocialQ&A compared to the prior research. The following chapter will provide insight into the design of the SocialQ&A system.

CHAPTER 3

SOCIAL Q&A UNDER THE HOOD

This chapter describes in detail the design of SocialQ&A. First, it briefly introduces the components of SocialQ&A and describes the high-level functionality of each of these components. Second, the flow of events in SocialQ&A is introduced to explain the methods employed by SocialQ&A. Finally, each component and their interactions are described in detail.

3.1 Architecture of SocialQ&A

The objective of this research is to design a Q&A system to improve the quality of answers and decrease wait times by leveraging social networks. Thus, we developed algorithms to leverage the aspects of social networks and implemented a real-world social network-based Q&A system, called SocialQ&A, that utilizes user profile information, user action history, and user interactions in the social network. A detailed description of the core components of SocialQ&A is presented in this section and the algorithms used to realize the functions of each component are provided in Section 3.3.

Figure 3 shows the high-level architecture of SocialQ&A and the interaction between the core components: (1) User Interest Analyzer, (2) Question Categorizer, and (3) Question-User Mapper. Component (1) analyzes data associated with each user in the social network to derive user interests. Component (2) categorizes the end user questions into an interest category based on the category Synsets from WordNet. Based on information from Component (1) and Component (2), Component (3) forwards the questions from the end user to users who are likely able to satisfactorily answer the questions. The data from end user questions and subsequent answers is stored on a server to serve subsequent similar questions.

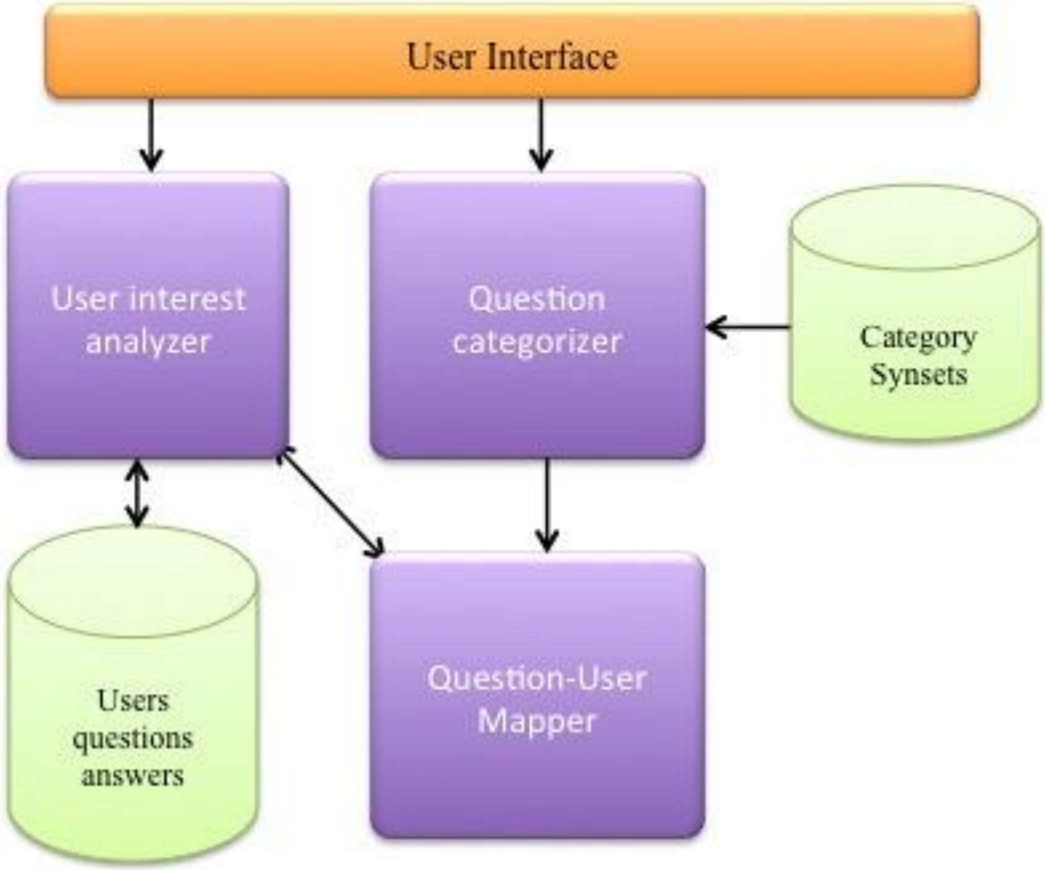


Figure 3. Architecture of SocialQ&A and the interaction between core components.

SocialQ&A is implemented using the Model-View-Controller (MVC) architecture [51]. The MVC architecture is widely adopted by software developers and is one of the most common software engineering architectures. The primary motivation for using the MVC architecture is the separation of concerns as mentioned by Krasner *et al.* [51].

3.1.1 *User Interest Analyzer*

The *User Interest Analyzer* utilizes data derived from the user's profile information and user interactions (questions asked and answers provided) in the social network to determine the interests of the user more accurately in terms of various pre-defined interest categories. A total of 36 pre-defined interest categories, including sub-categories derived from the Yahoo! Answers Q&A system were used to implement SocialQ&A. Examples of the major categories include music, movies, television, and books.

It is straightforward to derive a user's interests directly from the interest list in his/her profile. Tracking user interactions in the system to derive user interests is accomplished by using the tags related to questions either asked or answered by the user. In this way, SocialQ&A updates the user's interests regularly. The intuitive reason behind such a design is that if an end user asks a question, the question categories indicate that the end user is interested in those particular categories. The dynamic interaction tracking implemented in SocialQ&A for interest derivation provides a more accurate reflection of user interests than the static approach that depends solely on the user's profile information to represent user interests.

The derived interests of each user are represented by a user-interest vector. Figure 4 shows an example of a user-interest vector. The top line shows the pre-defined interest categories in the system and each column indicates an interest. In the figure, the value 0 indicates that user X does not have the corresponding interest, while the value 1 indicates that the user has the corresponding interest. Thus, each user is associated with a user-interest vector indicating his/her interests.

	Rock	Classic	Action	Thriller	News	Shows	Story
User X	1	0	1	0	0	0	1

Figure 4. User-interest vector.

3.1.2 *Question Categorizer*

The primary task of the *Question Categorization Component* is to categorize a question into a pre-defined interest category based on the topic of the question. The same pre-defined categories introduced above are used to categorize/tag the questions. The *Question Categorization Component* takes into consideration the tags (which are the same as the pre-defined categories) provided by the end user to categorize the question.

In addition to these tags, SocialQ&A uses WordNet to examine the text of the question and generate a stream of tokens by parsing the question string. These tokens are compared to the Synset that is created from the predefined categories to determine the category or categories where the question belongs. This process aims to categorize the question more accurately, taking into account that the user may omit some tags, tag inaccurately, or not tag the question at all.

3.1.3 *Question-User Mapper*

The *Question-User Mapper* performs the important task of utilizing the gathered information to identify the appropriate answer provider. To map a question to an answer provider, two parameters are considered: (1) Interest of the potential answer provider in the question topic(s), and (2) The social connectedness between the potential answer provider and the end user. After creating a list of potential answer providers, the *Question-User Mapper* sorts them based on the probability of being able to answer the question and dispatches the list of top answer providers to the *Notifier*. The *Notifier* is responsible for notifying the potential answer providers in the list.

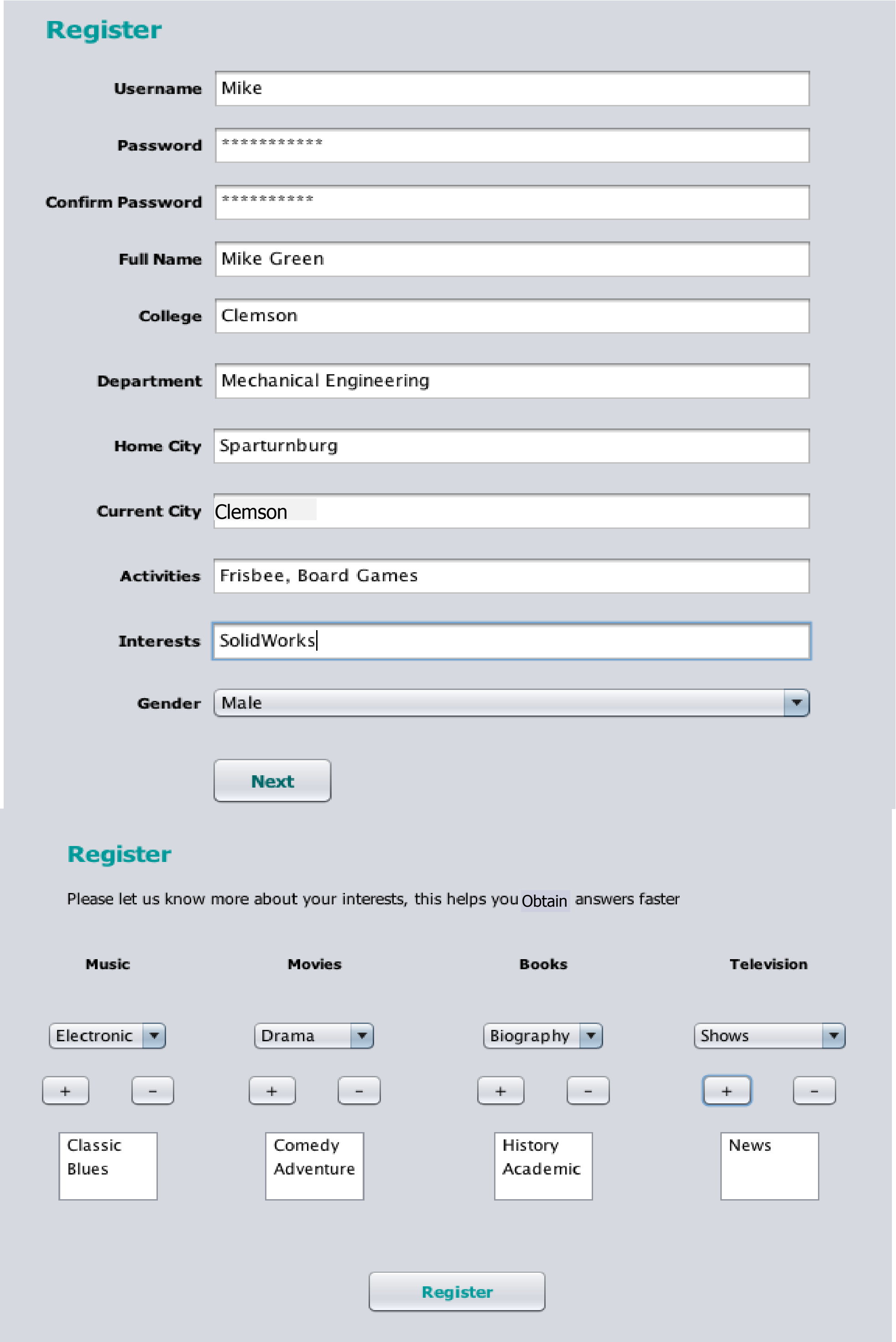
3.2 Flow of events

The user's interactions with the system can be performed on two fronts: the Q&A domain and the social platform. The goal of the system is to make efficient use of user interactions on both of these fronts to improve the user experience and satisfaction in the Q&A system.

Consider a hypothetical user of the system named Mike. When Mike registers for SocialQ&A, he is required to provide essential information about himself, such as his personal information, area of study/expertise, his current interests, and his involvement in other activities. Users are also encouraged to describe their interests in terms of a few pre-defined categories shown in the screenshots of the registration views (Figure 5). SocialQ&A uses the registration information to determine Mike's expertise/interest in particular topics. SocialQ&A then uses the interest information to determine how closely Mike's interests match the question topics. If Mike's interests match the question topics, he is identified as a potential answer provider for the question.

When a user logs in, he/she is prompted to add friends to build or expand his/her current social network. The formation of a broad social network is an important aspect of SocialQ&A. When a user adds a friend, in addition to constructing the social links, SocialQ&A also determines the similarity of interests among the friends.

Interest similarity is taken into account when determining the list of answer providers to whom the question could be routed. Interest similarity between two users is calculated using the Hamming distance between the interest vectors of those users. To calculate the Hamming distance, the interest vectors of the users are compared to each other one element at a time; when two elements at the corresponding positions are the same, the count for the Hamming distance is incremented.



Register

Username	Mike
Password	***********
Confirm Password	**********
Full Name	Mike Green
College	Clemson
Department	Mechanical Engineering
Home City	Sparturnburg
Current City	Clemson
Activities	Frisbee, Board Games
Interests	SolidWorks
Gender	Male

Register

Please let us know more about your interests, this helps you Obtain answers faster

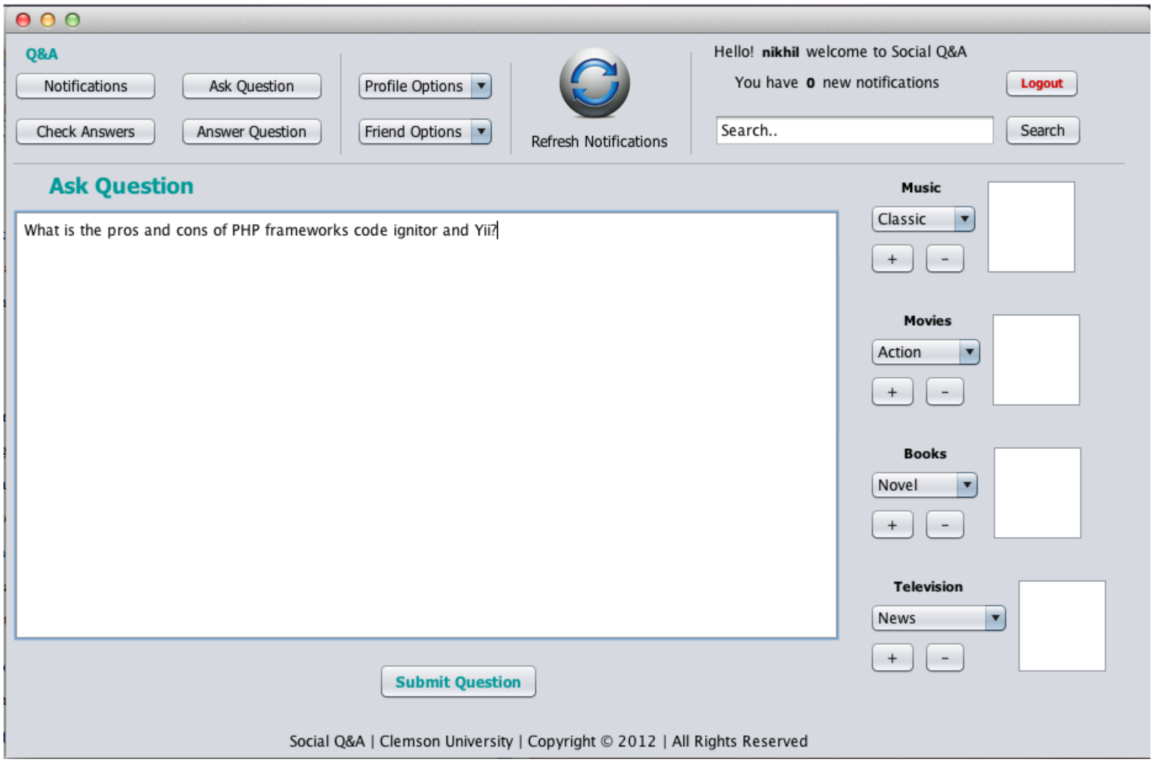
Music	Movies	Books	Television
Electronic	Drama	Biography	Shows
Classic Blues	Comedy Adventure	History Academic	News

Figure 5. Registration example.

The rationale behind this approach is that when an answer provider knows the end user who posted the question and they have many interests in common, he/she is more motivated to answer the question than if they are strangers or have few common interests [62].

Another feature provided by SocialQ&A is the option to forward questions. In the earlier example, suppose Mike is notified of a question posted by one of his friends. Mike himself is not capable of answering that particular question, but he has a friend in his social network who he believes would be able to provide an answer to the posted question. In such a situation, Mike can personally forward the question to his friend.

Another significant chain of events is set in motion when an end user posts a question. Figure 6 is a screenshot of the end user's view for asking questions. The end user is allowed to tag a question based on his/her perception of the interest category of that question.



Hello! **nikhil** welcome to Social Q&A

You have **0** new notifications

Search..

Ask Question

What is the pros and cons of PHP frameworks code ignitor and Yii?

Music Classic

Movies Action

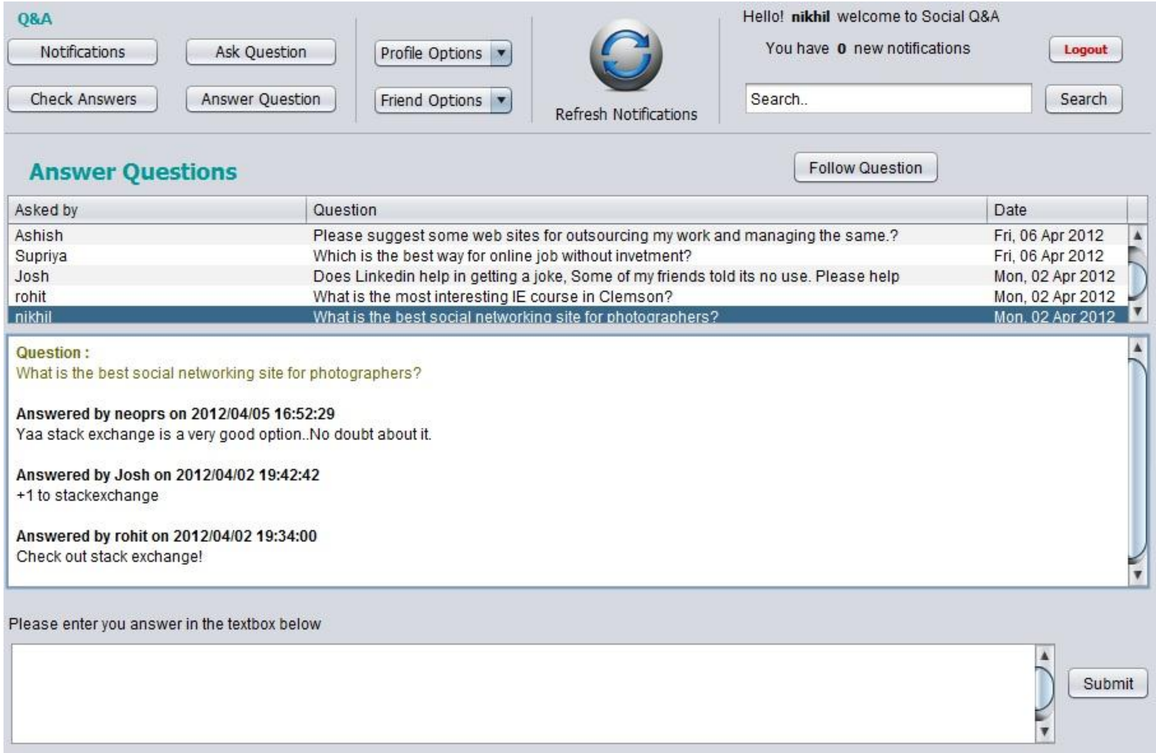
Books Novel

Television News

Social Q&A | Clemson University | Copyright © 2012 | All Rights Reserved

Figure 6. User view for asking a question.

Based on the aforementioned characteristics including the social network of the end user, the tags provided by the end user, and the tags assigned by the Question Categorizer, Social Q&A determines potential answer providers and routes the question to those providers. The social network of the end user is used to determine his/her friends and how closely their interests match with the end user. The tags provided by the end user and those assigned by the *Question Categorizer* Component are used to determine whether the potential answer providers have interests matching the question category. If no user is able to answer a question, then the end user who posted the question would never receive an answer for the question. This limitation exists in all Q&A systems. An example question and answer thread is shown in Figure 7.



Hello! **nikhil** welcome to Social Q&A

You have **0** new notifications

Answer Questions

Asked by	Question	Date
Ashish	Please suggest some web sites for outsourcing my work and managing the same.?	Fri, 06 Apr 2012
Supriya	Which is the best way for online job without invetment?	Fri, 06 Apr 2012
Josh	Does Linkedin help in getting a joke, Some of my friends told its no use. Please help	Mon, 02 Apr 2012
rohit	What is the most interesting IE course in Clemson?	Mon, 02 Apr 2012
nikhil	What is the best social networking site for photographers?	Mon, 02 Apr 2012

Question :
What is the best social networking site for photographers?

Answered by neoprs on 2012/04/05 16:52:29
Yaa stack exchange is a very good option..No doubt about it.

Answered by Josh on 2012/04/02 19:42:42
+1 to stackexchange

Answered by rohit on 2012/04/02 19:34:00
Check out stack exchange!

Please enter you answer in the textbox below

Figure 7. An example of a question and answer thread.

Unlike previous Q&A approaches, SocialQ&A exploits the users' profile information and interests, in addition to the end user's social network and Q&A activities to determine potential answer providers. Additionally, the interest information of all users in the system is continuously updated based on their actions. SocialQ&A also differs from other Q&A approaches by routing questions only to potential answer providers, thereby reducing the number of notifications sent to users. However, any user can still see and potentially answer all the questions asked by any end user of SocialQ&A.

3.3 Core algorithms

This section provides a detailed description of the three core algorithms that drive SocialQ&A: (1) *User Interest Analyzer,* (2) *Question Categorizer* and, (3) *Question-User Mapper.* These algorithms are used to analyze user information, sort questions, and determine potential answer providers, respectively.

3.3.1 User Interest Analyzer

The main purpose of the *User Interest Analyzer* is to map users to their interests. Figure 8 is a depiction of the process flow, and pseudocode is provided in Algorithm 1. As the left side of the figure shows, whenever a user registers for a new account, a data entry is created for that account in the database. The end user is then presented with the home page, so that he/she can continue his/her activity. The User Interest Analyzer algorithm (the right part of the figure) is executed in a separate thread (Algorithm 1).

When a user registers, he/she is given the option of entering his/her interests and activities in text and to choose from pre-defined interest categories to add to his/her interest list, as shown in Figure 8. These text fields are then parsed to generate token streams (Steps 1,2,3). For every token in a given token stream, its matching interest category is located in the Synset (Step 4).

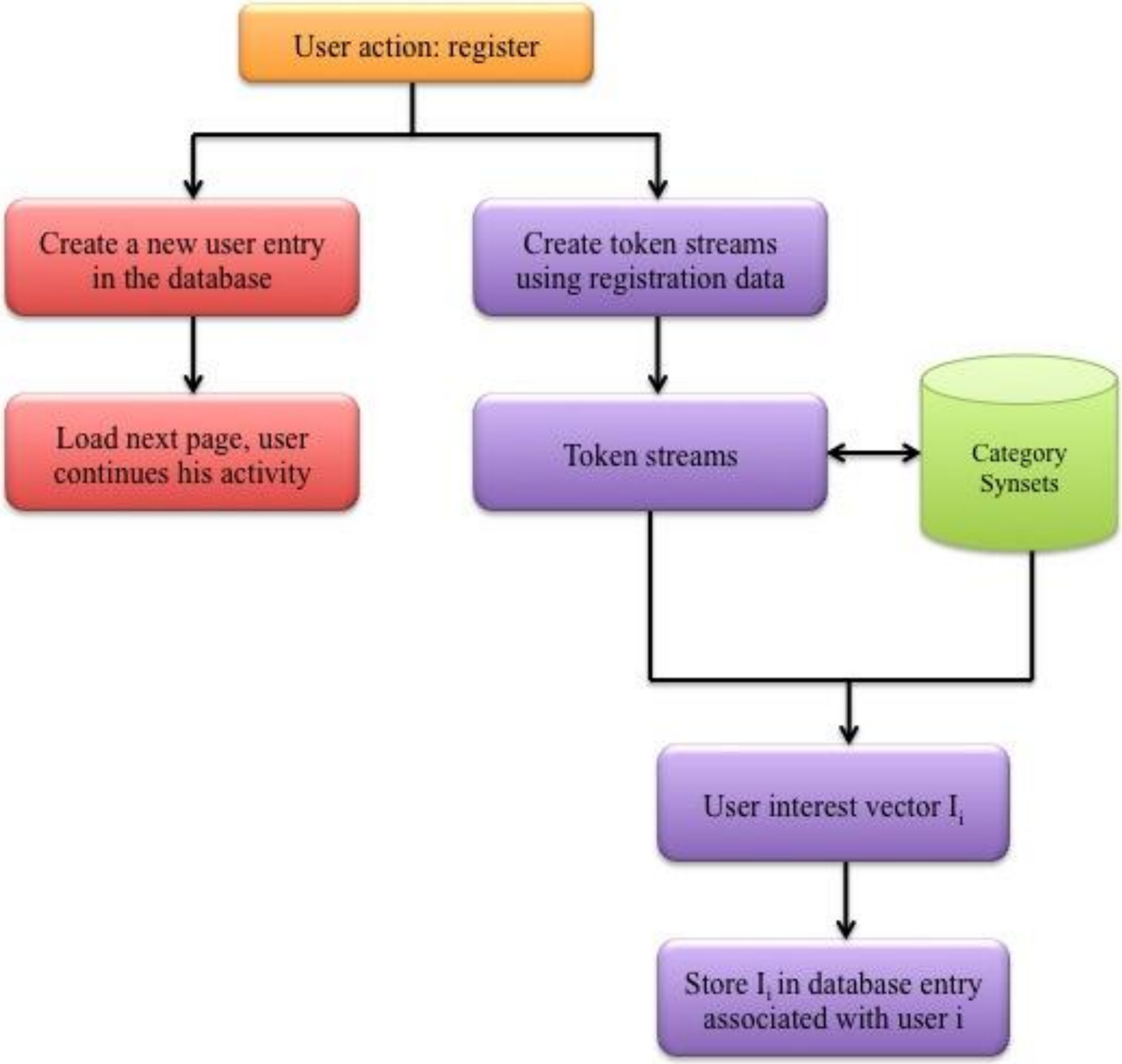

User action: register
Create a new user entry in the database
Create token streams using registration data
Load next page, user continues his activity
Token streams
Category Synsets
User interest vector I_i
Store I_i in database entry associated with user i

Figure 8. Representation of the *User Interest Analyzer* algorithm.

Finally, an interest vector is generated for that user. Since the *User Interest Analyzer* algorithm requires significant computation time, it is encapsulated inside an asynchronous thread to ensure that it does not interfere with user actions.

Algorithm 1: Pseudocode for the User Interest Analyzer algorithm.

Input: *A user's profile*
Output: *A user's interest vector*

```
Start:
    1. Parse the "interests" field in the user's profile to generate a
       stream of tokens $T_i$ .
    2. Parse the "activities" field in the user's profile to generate a
       stream of tokens $T_{a.}$
    3. Use the inputs from the user's selection from the Music,
       Movies, Television and Books fields in the user's profile to
       generate streams of tokens $T_{mu}$, $T_{mo}$, $T_t$ and $T_b$.
    4. For every token stream $T_x$,($T_x$ is $T_{mu}$, $T_{mo}$,$T_t$ or $T_b$)
           a. Compare each token to the Synset of pre-defined
              categories.
           b. If a matching interest category of the token exists in
              the Synset, add that category to the user's interest
              vector 'I'. For example, if the category music is
              matched, I[music] = 1.
    5. Store Vector I in the database as the user's interest.
End.
```

After the algorithm completes, the user is associated with a vector of interests. Figure 9 shows an example of the User-interest matrix. The database consists of a 2-dimensional matrix of size *m* x *n*, where *m* corresponds to the number of users and *n* corresponds to the number of pre-defined categories. The numbers in the figure represent the weights of that interest for each user. The weight represents the degree of a user's interest in a category. For example, if a user (user 5) has asked/answered a lot of questions regarding the rock category, then the number in the rock field will be higher. Weight calculation will be explained in detail in a later section.

	Rock	Classic	Action	Thriller	News	Shows	Story
User1	5	8	2	2	1	8	4
User2	3	3	7	5	8	3	6
User3	0	7	5	9	2	6	8
User4	8	8	1	0	7	2	2
User5	12	6	0	2	1	0	8
User6	6	0	0	5	0	0	1
User7	3	5	12	8	1	1	1

Figure 9. User-interest matrix visualization.

3.3.2 Question Categorizer

The *Question Categorizer* algorithm categorizes a given question in terms of predefined categories. Analogous to the *User Interest Analyzer*, the *Question Categorizer* strives to associate a vector R_i to a given question Q_i, where R_i is the vector of predefined categories corresponding to question Q_i. The format of a question vector is the same as in Figure 4. Algorithm 2 shows the pseudocode for categorizing a question.

When an end user posts a question as shown in Figure 6, he/she can choose tags in the categories movies, music, books, and television for the question (Step 1). The question is then parsed to generate a token stream (Step 2). For every token in the token stream, its matching interest category is located in the Synset (Step 3). Finally, an interest vector is generated for the question.

The questions posted by a user are used to dynamically update his/her interest vector and interest weights. The interest weight in a user's interest vector represents his/her degree of interest and is used to more accurately reflect the user's interests. The interest weights in the vector generated during registration (the categories indicated by the user) are initialized to one. Later, each time a user asks a question, the question is parsed to a question vector using the method previously explained. As shown in step 4 of Algorithm 2, SocialQ&A checks whether each element in the question vector exists in the user's interest vector. If yes, the weight of this element in the interest vector is incremented by one. Otherwise, this element is added to the interest vector with an initial weight of one. For example, if a user asks a question in the "movies" category and his/her interest vector includes "movies", then the weight of the interest category "movies" is incremented. If a question belongs to two or more categories, the weights of the multiple corresponding interest categories are incremented. Therefore, a user's interest vector always reflects his/her most recent interests. The rationale for this method is that if an end user is asking a question belonging to a certain category ‘x’, then he/she has an interest in that category even though he/she did not indicate it while creating his/her profile. This method can be extended by considering the questions answered by the user. This weight adjustment serves to dynamically update each user’s interest information. Thus, the system is gradually learning more and more about a user every time he/she performs a Q&A activity; this improves the question routing performance of the system. Steps 4-6 in Algorithm 2 show the pseudocode for the dynamic interest adjustment.

The *Question Categorizer* algorithm associates each question Q_i with a vector R_i, which results in a 2-dimensional matrix representation of size *m* x *n*, similar to the user-interest representation. The only change is *m* corresponds to the number of questions. Figure 10 depicts the process flow. After the *Question Categorizer* algorithm completes, it delegates control to the *Question-User Mapper* to determine a list of potential answer providers.

Algorithm 2: Pseudocode of the Question Categorizer algorithm.

Input: *A question posted by a user*

Output: *A question vector and an updated interest vector*

```
Start:
  1. Initialize the question vector 'R' with the tags indicated by the end user.
  2. Parse the question to generate a stream of tokens $T_q$.
  3. For every token in the token stream $T_q$,
       a. Compare the token to Synset of pre-defined categories.
       b. If the entry of the token exists in Synset, add the mapping interest category to vector 'R'. For example, for category computer, R[computer] = 1 if it is zero.
  4. For each element in vector 'R', check whether it exists in the end user interest vector I (with interest weight denoted IW),
       a. If yes, increment the weight associated with that entry. For example, if $I_i$ [thriller] == 1, $IW_i$ [thriller]++ ;
       b. If no, add the element to the interest vector. For example, if $I_i$ [thriller] == 0, $I_i$ [thriller] == 1 ;
  5. Store the Vector R in the database along with the question.
  6. Pass control to Question-User Mapper to find potential answer providers.
End.
```

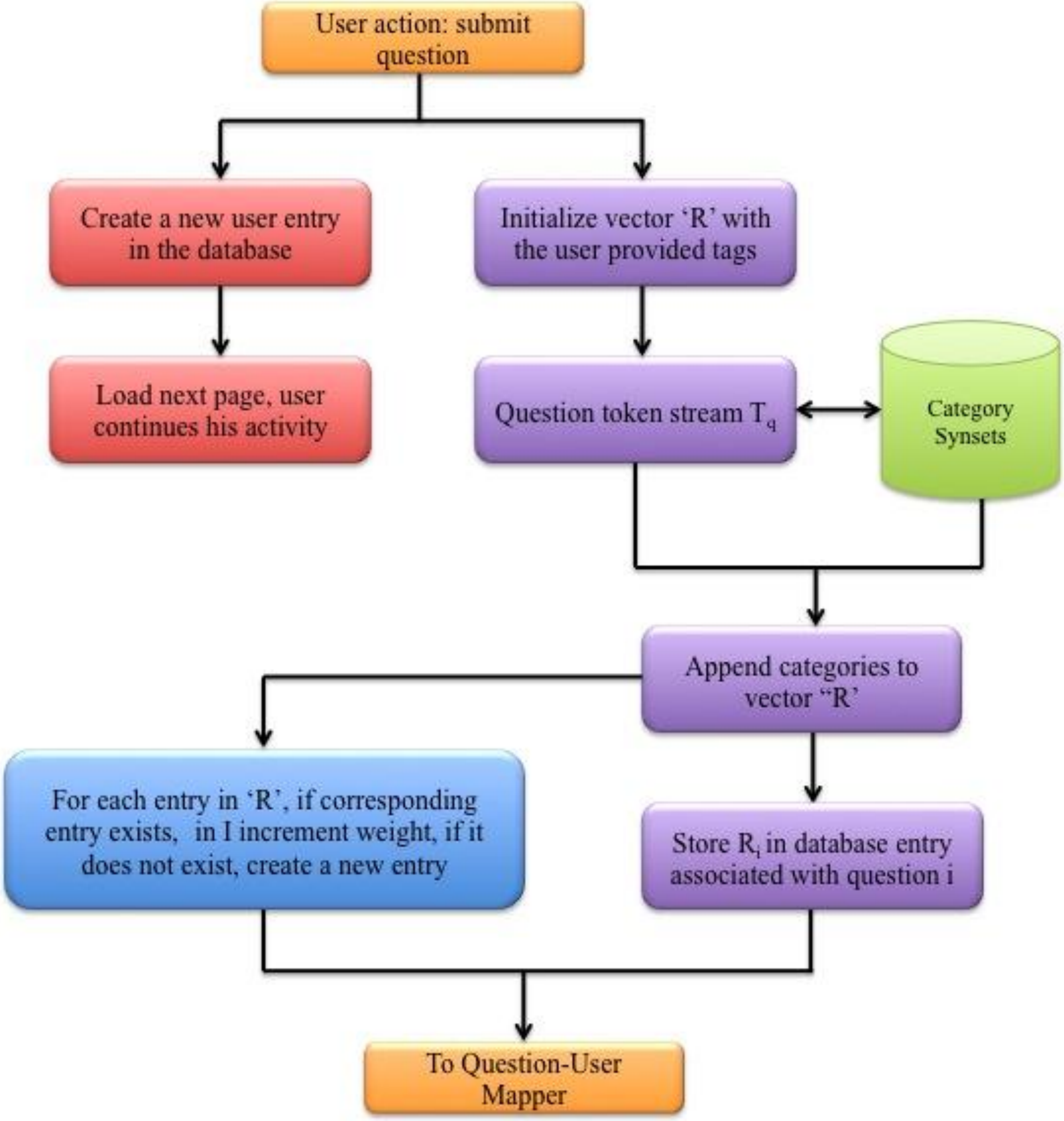

User action: submit question
Create a new user entry in the database
Initialize vector 'R' with the user provided tags
Load next page, user continues his activity
Question token stream T_q
Category Synsets
Append categories to vector "R'
For each entry in 'R', if corresponding entry exists, in I increment weight, if it does not exist, create a new entry
Store R_i in database entry associated with question i
To Question-User Mapper

Figure 10. Representation of the *Question Categorizer* algorithm.

3.3.3 Question-User Mapper

The *Question-User Mapper* algorithm is the central focus of this research. The chief task of the *Question-User Mapper* algorithm is to consider both the interests of potential answer providers in the categories of the question and the social connectedness between the potential answer providers and the end user to generate a list of potential answer providers with the ability to provide a satisfactory answer. Then, it sorts the list based on the ability to answer the question and forwards the question to the users on that list. The question is forwarded to the top answer providers, i.e. those with the highest metrics in the list. The flow of the algorithm is provided in Figure 11 and the pseudocode is provided in Algorithm 3.

While computing the list, SocialQ&A considers two factors in the process of selecting the optimal list of potential answer providers:

1. The interest of a potential answer provider in the categories of the question (the user interest factor *Fi*).
2. The social connectedness between the potential answer provider and the end user (the social connectedness factor *Fc*).

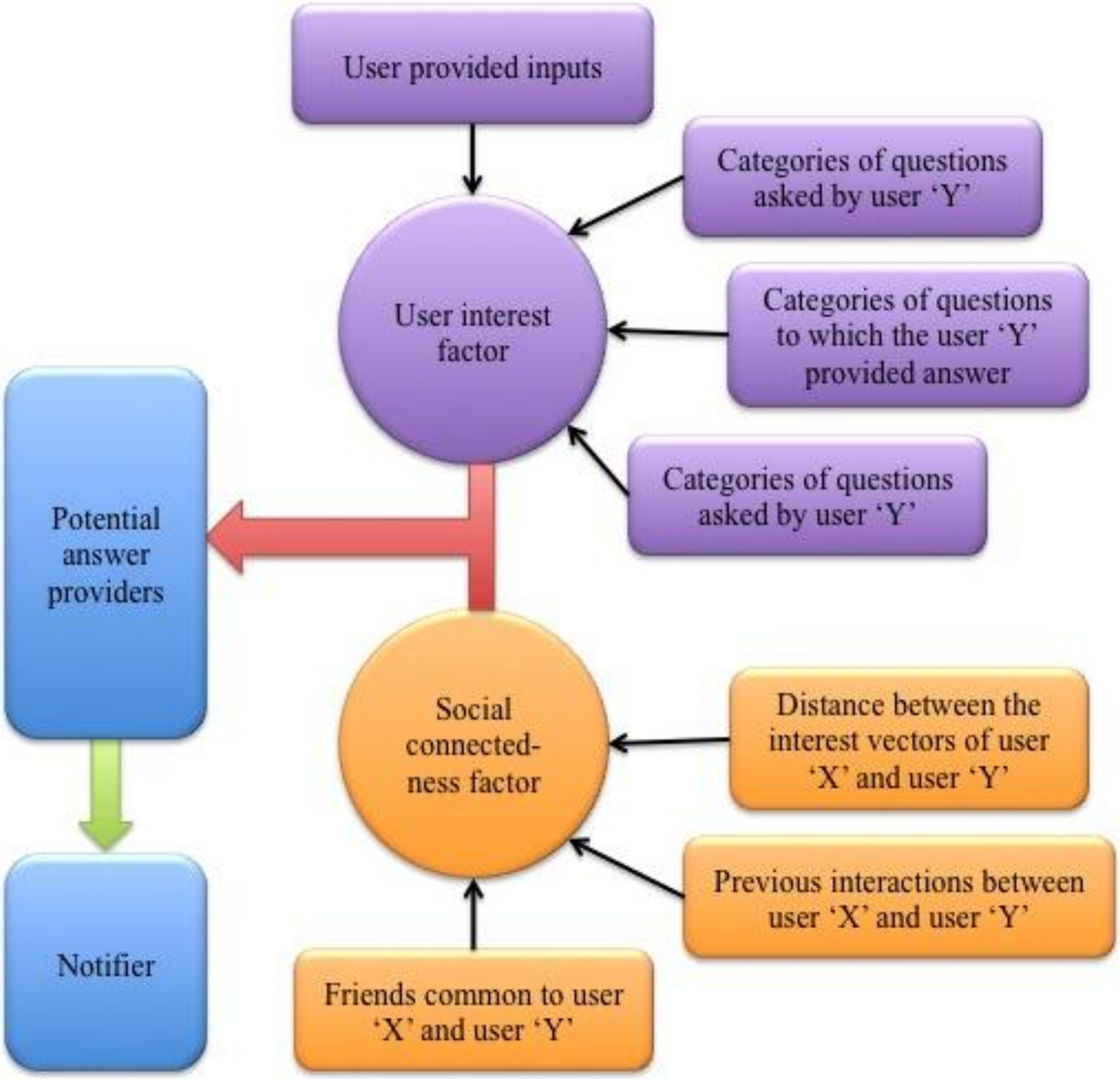

User provided inputs
Categories of questions asked by user 'Y'
User interest factor
Categories of questions to which the user 'Y' provided answer
Categories of questions asked by user 'Y'
Potential answer providers
Social connected-ness factor
Distance between the interest vectors of user 'X' and user 'Y'
Notifier
Previous interactions between user 'X' and user 'Y'
Friends common to user 'X' and user 'Y'

Figure 11. Representation of the *Question-User Mapper* algorithm.

Algorithm 3: Pseudocode of the Question-User Mapping algorithm

```
Input: Interest vectors of the end user and his/her friends and a
question vector
Output:  A list of potential answer provider

    Start:
        1. Let $R_q$ be the vector of categories to which the question q
           belongs

   For each friend (y) of the end user (x) do the following
        2. Let $I_y$ be the interest vector of user 'y'.
            //Calculate the user interest factor(steps 3-4)
        3. Compute the common interests between the vectors $R_q$ and $I_y$
           and calculate Fi.
        4. Let $I_x$ be the interest vector of user 'x'.
            //Calculate the interest similarity between 'x' and 'y' (step 5)
        5. Compute the Hamming distance ($D_i$) between $I_x$ and $I_y$ as IS.
           //Calculate the interactions between 'x' and 'y' (step 6)
        6. Let 'n' be the number of previous interactions between user
           'x' and user 'y' (PI).
            //Calculate the number of common friends between 'x' and
           'y' (step 7)
        7. Let $C_f$ be the number of friends common to both user 'x' and
           user 'y' (CF).
        8. Calculate the final metric for the end user's friends using the
           equation:"$F = 2Fi + Fc$" ($Fc = IS + PI + CF$).
        9. Order the friends by the final metric in descending order.
        10. Create a list containing the top 'k' friends.
        11. Present this list to the notifier to notify the appropriate users.

    End.
```

The user interest factor *Fi* of a potential answer provider to a question is calculated from the composition of four elements:

1. The interest categories of the potential answer provider derived from the data provided by the user during registration (*Ui*);
2. The interest categories mined from the questions asked by the potential answer provider (*Uq*);
3. The interest categories associated with the questions asked by other users to whom the potential answer provider under consideration has provided an answer (*Ua*);
4. The interest categories associated with the question calculated using question categorizer algorithm (*Rq*).

The first three elements are actually the elements used for determining the interest vector of a user. These four elements are combined using *Equation 1* to find the common interest categories (Fi) between the potential answer provider and the question.

Equation 1: $Fi = Rq \cap (Sum(Ui) \cup Sum(Uq) \cup Sum(Ua))$

The first element (*Rq*) in *Equation 1* is a vector of categories of a question. The second element (*Ui*) in *Equation 1 is* an interest vector containing the interests of a user as shown in Figure 4. The vector initially consists of interests entered by the user during registration. Subsequently, when the user asks a question or answers a question, the third (*Uq*) and fourth (*Ua*) elements in *Equation 1* are updated, and the interest categories of the question are added to the end user’s interest vector. If the categories already exist, the weights of the corresponding entries in the vector are incremented. Finally, the interests in the vector along with their weights, represent the user’s interests.

For computing the social connectedness factor *Fc* between a potential answer provider and an end user, we consider the following:

1. The similarity between the interest vectors of the potential answer provider and the end user (IS);
2. The interactions between the potential answer provider and the end user, e.g., the number of questions asked by user ‘x’ and answered by user ‘y’ and the number of questions asked by user ‘y’ and answered by user ‘x’ (PI);
3. The number of common friends between the potential answer provider and the end user (*CF*).

Using these metrics, the system determines a social connectedness factor *Fc* that increases with the similarity between interest vectors, number of interactions, and number of common friends between the potential answer provider and the end user, as shown in *Equation 2*.

Equation 2: $Fc = IS + PI + CF$

To calculate the interest similarity *IS* in *Equation 2*, we match the interest vectors of the two users. Each matching entry in the two interest vectors increments the value of interest similarity by one. To calculate the interaction element *PI* in *Equation 2*, we determine the number of questions asked by user 'x' and answered by user 'y,' and vice versa. The third element *CF* in *Equation 2* is simply the number of friends common to user 'x' and user 'y'. The sum of all three elements gives the social connectedness factor *Fc*.

The final list of potential answer providers is determined by considering both factors described above (user interest factor *Fi* and social connectedness factor *Fc*). The user interest factor *Fi* represents the potential ability of a user to answer the question, and the social connectedness factor *Fc* represents the willingness of a user to answer the question. *Equation 3* is used to calculate the final metric *F*, where *Fi* is multiplied by '$\propto$', and the social connectedness factor *Fc* is multiplied by '$1-\propto$'. Parameter $\propto$ denotes the consideration weight for each parameter, and it enables the system to set different priorities for *Fi* and *Fc* based on their influences on identifying appropriate potential answer providers.

Equation 3: $F = \propto Fi + (1-\propto)Fc$

Studying the influence of the two factors and deterministic calculation of '$\propto$' is a non-trivial task, which would require repeated experiments of the real-world system using different values of '$\propto$'. Thus, this task remains as future work. Since *Fi* should have a higher influence than *Fc* intuitively, for the current implementation, we have set '$\propto$' to 0.67. Thus, the contribution of the user interest factor is twice that of the social connectedness factor in the calculation of the final metric.

As soon as the algorithm completes, the top potential answer providers as determined by the algorithm receive a notification for the posted question. Resembling the *User Interest Analyzer* and the *Question Categorizer* algorithms, the *Question-User Mapping* algorithm is also implemented as an asynchronous thread so that it does not interfere with other user actions. Thus, all three algorithms work together with the user-friendly front-end to make SocialQ&A an efficient and improved Q&A system.

3.4 Summary

In summary, this chapter provides a detailed description of the important components of SocialQ&A and the interactions between them. It also describes the algorithms developed as a part of this research. The next chapter discusses the results that were obtained by analyzing the data from our prototyped real-world SocialQ&A system.

CHAPTER 4

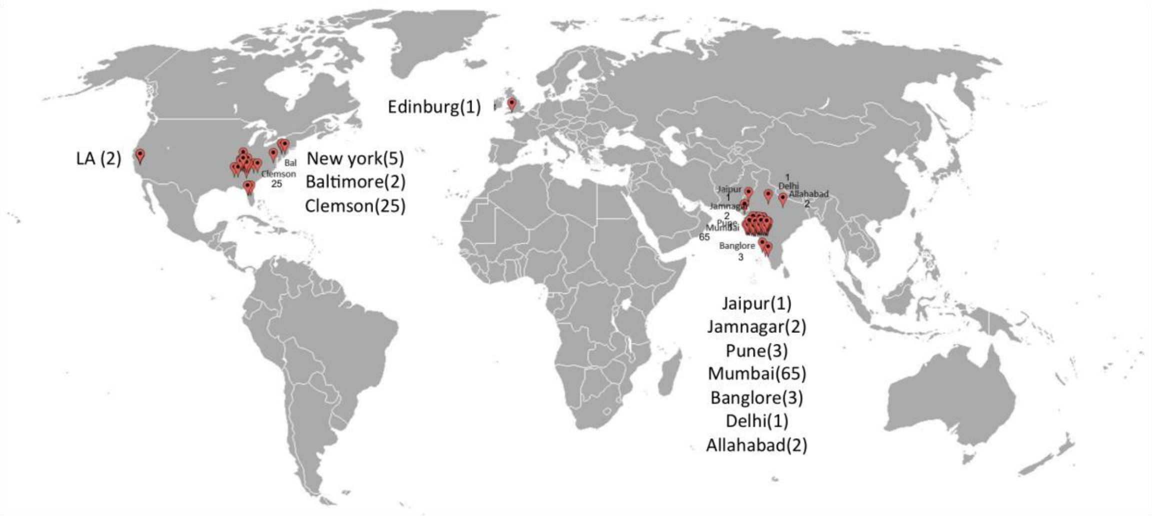
RESULTS AND ANALYSIS

This chapter presents the results and analysis based on the usage of SocialQ&A over a period of approximately one month beginning March, 2012. SocialQ&A was released to a limited group of individuals for experimental purposes. Over the one-month period, a total of 124 people registered and used SocialQ&A. 163 questions were posted and 282 answers were posted in response. For research purposes, these users were considered to be part of one social network. We requested the users to be online during certain time slots, at their convenience in order to have enough users online in the testing. The distribution of the 124 users in SocialQ&A is shown in Figure 12. Approximately 35 users were from the United States, 70 users were from India, and 1 user was from the United Kingdom.

In this research we have made the following assumptions:

1. Due to the limited number of registered users, we placed all users of the system in a single social network to better represent a relatively large individual social network of a user in practice (in typical social networks, users have hundreds of connections). Practically, it is very difficult to test the prototype system with millions of users to directly compare it with existing systems such as Yahoo! Answers.

2. We assume that the answer ratings in SocialQ&A increase as the number of users increases with the further assumption that the expertise in each question topic will also increase accordingly; thus, higher-quality answers can be given.
3. We assume that the wait time decreases as the number of users in SocialQ&A increases with the further assumption that the number of users online at the same time also increases accordingly, effectively reducing the wait time.



4. Figure 12. Users in SocialQ&A.

4.1 User questioning and answering activity

We used the number of questions and answers posted to characterize user activity. According to the data, out of 124 users, 75 unique users posted at least one question, with the remaining users posting no questions. Moreover, out of 124 users, 81 unique users provided at least one answer, and the remaining users provided no answers. Out of 124 users, 26 users (approximately 20%) did not post or answer any questions. Consequently the remaining 80% were not passive and did contribute actively to SocialQ&A in some way.

Figure 13 is the graph for the number of questions asked by each user, ranging from 0 to a maximum of 10. Figure 14 displays the percentage of users who asked a given number of questions. As seen from the figures, approximately 56% of the users asked just one question, approximately 23% of the users asked two questions, approximately 10% of the users asked 3 questions, and the remaining 11% asked more than 3 questions. Thus, we can conclude that most of the users were fairly active, which implies that users are relatively active in the Q&A systems incorporated with a social network.

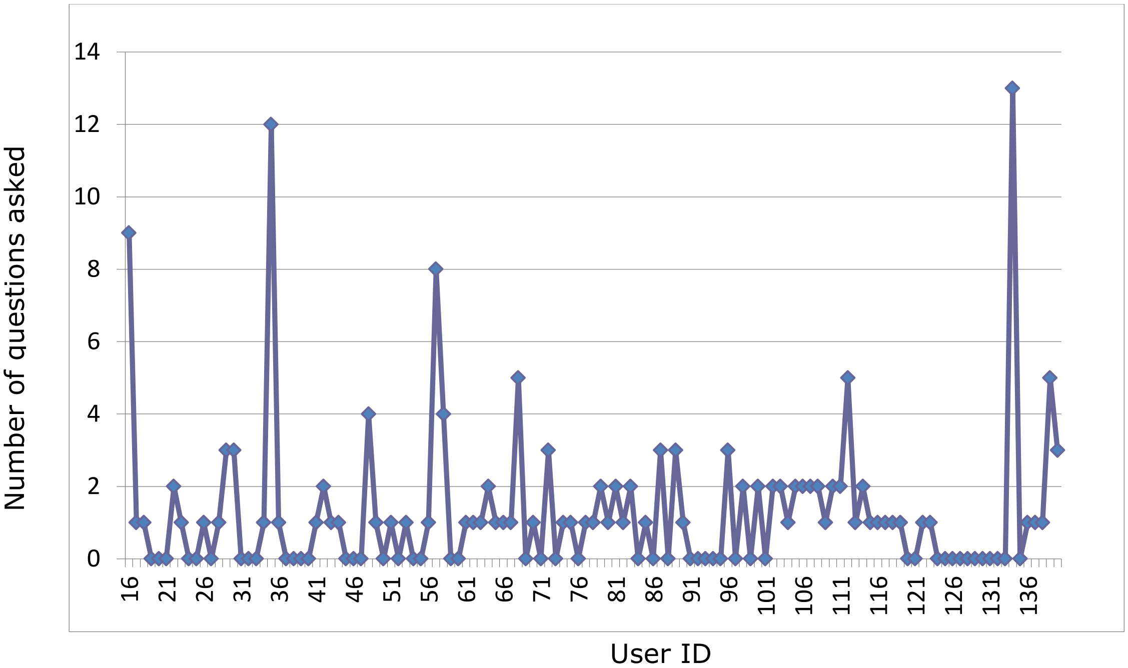


Figure 13. The number of questions asked of each user.

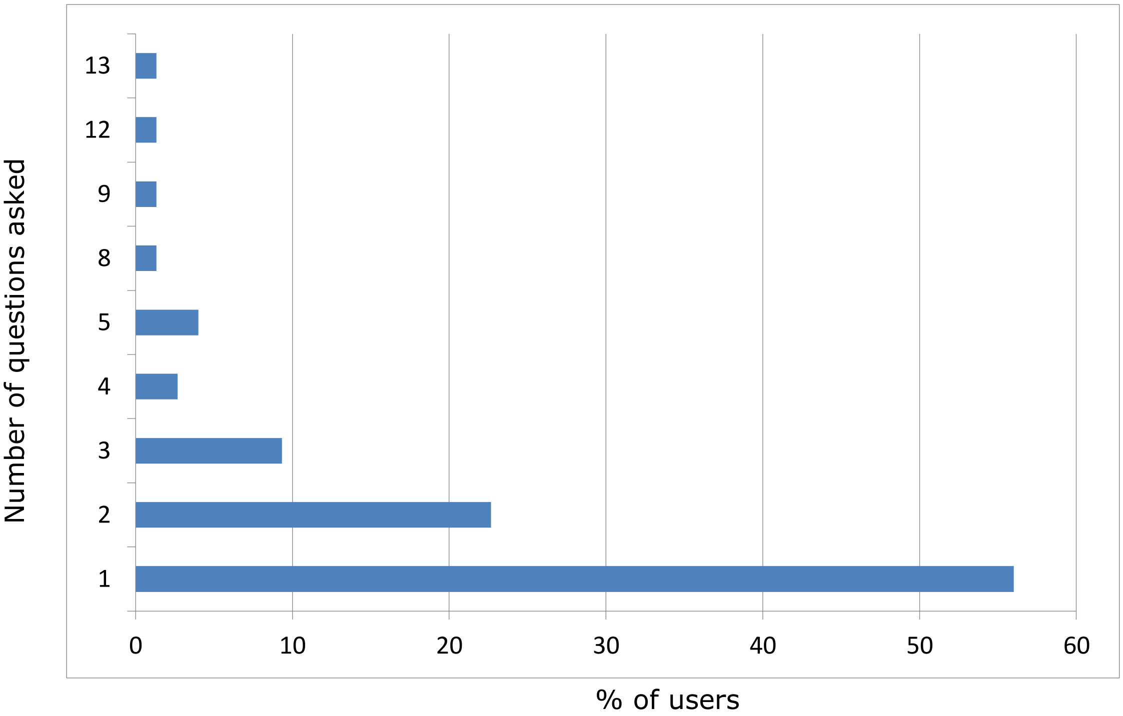


Figure 14. The number of questions asked vs. % of users.

Figure 15 shows the number of answers posted by each user, indicating the answering activity of the users. On average, users posted two to three answers. There are some users that were extremely active and posted five or more answers, and one of the users posted a total of 19 answers. Figure 16 shows the number of answers posted versus the percentage of users. Approximately 25% of the users provided just a single response, approximately 15% of the users provided 2 answers, 15% of the users provided 3 answers, approximately 10% of the users provided 4 answers, and approximately 40% of the users provided 4 or more answers. Therefore, comparing Figure 16 with Figure 14, we see that users in our study tend to answer questions more actively than they asked questions. The results show that the users are very willing to provide answers in SocialQ&A, which confirms that a social network can be leveraged to encourage users to answer questions.

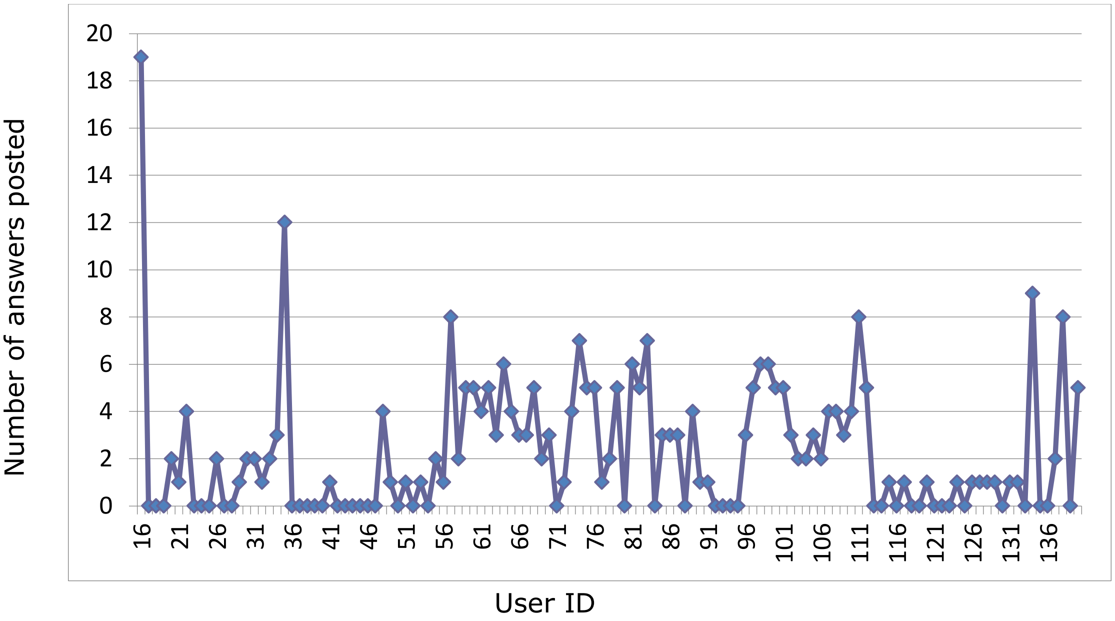


Figure 15. The number of answers posted by each user.

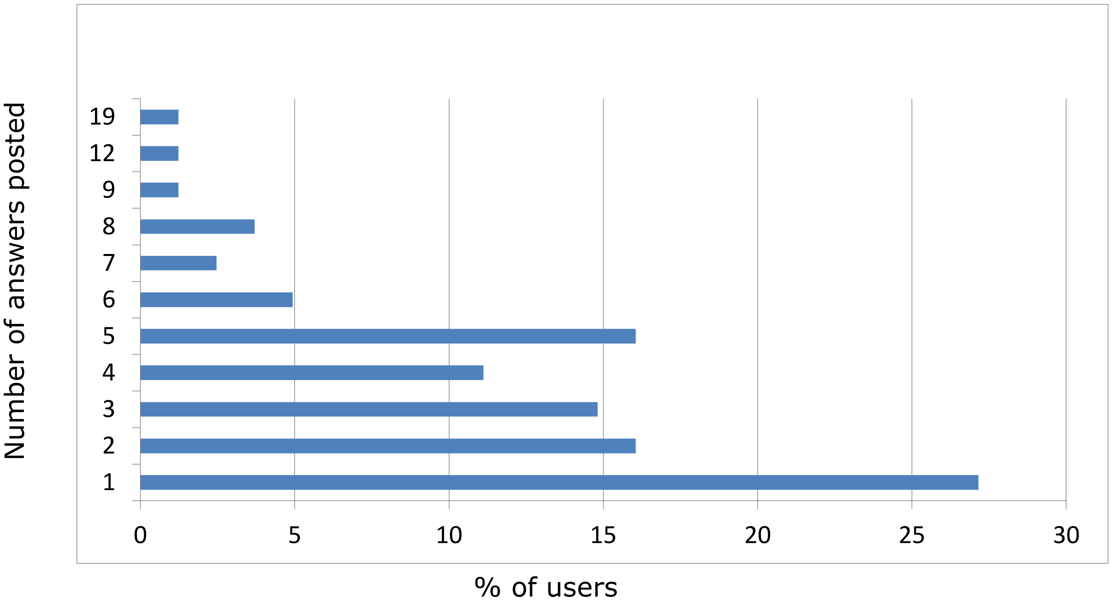


Figure 16. The number of answers posted vs. % of users.

As mentioned earlier, a total of 163 questions were asked and 282 answers were posted in response. A total of 24 out of 163 questions (around 15%) remain unanswered, while all other questions had at least one response. In comparison, Yahoo! Answers has more than 16% of questions unanswered, and Baidu Zhidao has 40% of questions unanswered [5]. Thus, at present, the percentage of unanswered questions in SocialQ&A is lower than those of Yahoo! Answers and Baidu Zhidao. As SocialQ&A identifies potential answer providers who have more common interests, close social relationships with the questioner, and interest in a question's category, those answer providers are more likely to answer their received questions. Thus, SocialQ&A is able to achieve an improvement even with a very limited number of users. Practically, we were not able to test SocialQ&A with millions of users. Hence, we do not claim that SocialQ&A is better than Yahoo! Answers or Baidu Zhidao. However, these results indicate a promising trend and that it is reasonable to assume that the system performance would increase as the number of users increases. We expect that the number of unanswered questions tends to reduce with an increase in users because with more users, the range of expertise also becomes broader, and the probability of a larger number of people being online at the same time a question is posted increases. Thus, SocialQ&A demonstrates its potential to improve on current Q&A systems.

Potential benefit of SocialQ&A*: The questions in SocialQ&A are more likely to be answered since the potential answer providers have a close social relationship with the end user and have an interest in the question category.*

4.2 Analysis of the questions

In this Section, we analyze the questions asked in SocialQ&A. The aspects analyzed are (1) Question paraphrasing, (2) Question categories, (3) Question types, and (4) The number of answers received per question. To determine the question types, categories and subcategories to which the question belongs, we manually examined every question.

As mentioned earlier, a total of 163 questions were posted. After analyzing those questions, we found that the average number of characters per question is 45.5 (10.65 words). The majority of questions (91%) are comprised of a single sentence. Approximately 75% of the questions were properly paraphrased with a question mark, although some questions contained multiple question marks.

As mentioned earlier, SocialQ&A uses four major categories: music, books, movies, and television. Figure 17 shows the distribution of questions among the four major categories. Approximately 38% of the questions were based on music, 29% were based on books, 41% were based on movies, and 13% were based on television. The percentages were calculated with respect to the total number of questions asked. Also, a question can belong to more than one category and such a question appears under all of its categories rather than just one. For example, in a total of 6 questions, 3 questions belong to category *x*, 1 question belongs to category *y,* and the remaining 2 questions belong to both category *x* and category *y,* then we say that approximately 83% (5 out of 6) of questions belong to category *x* and approximately 50% (3 out of 6) questions belong to category *y*. The 4 categories described earlier are further divided into a total of 32 subcategories. Figure 18 shows the distribution of questions among the various subcategories. These results indicate the interests of the current users in SocialQ&A.

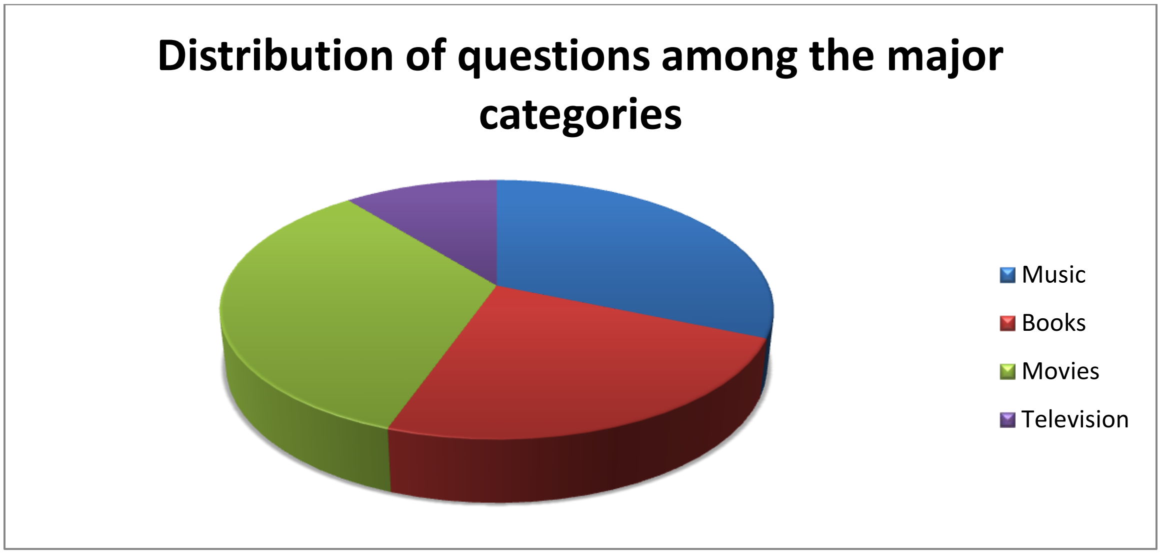


Figure 17. Distribution of questions among the major categories.

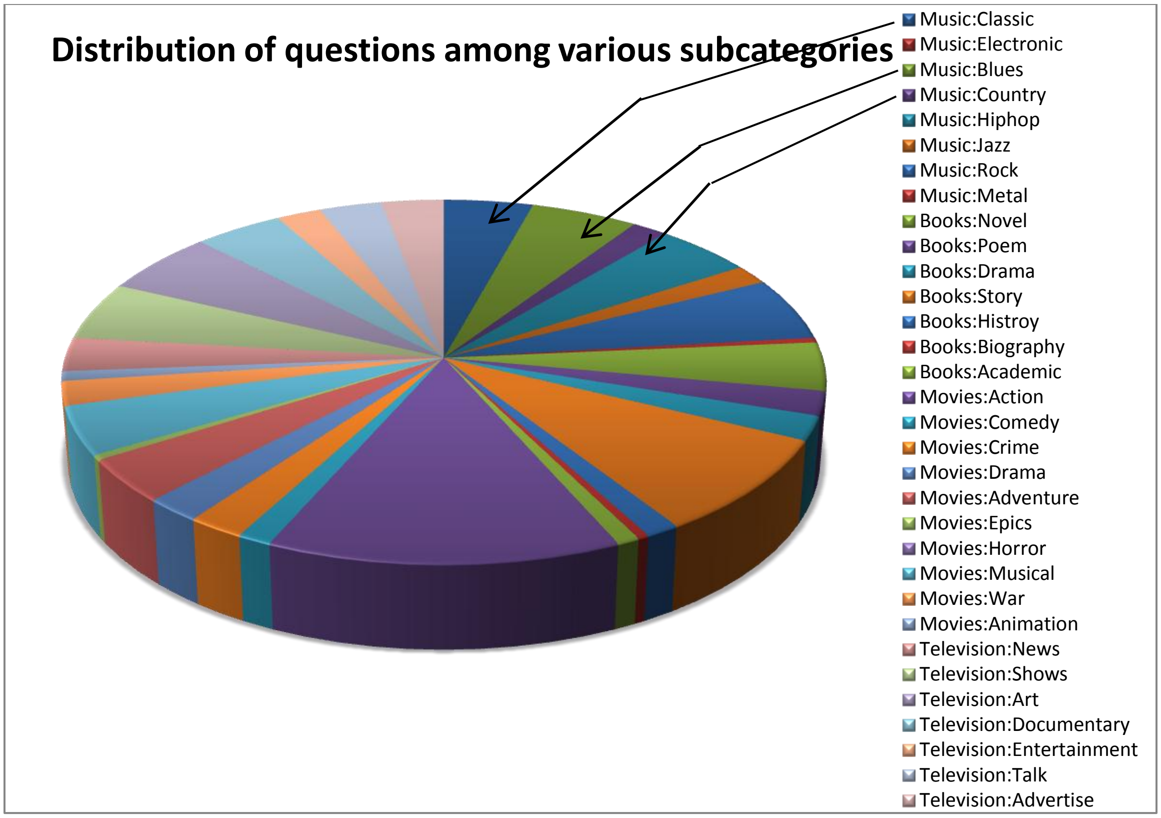


Figure 18. Distribution of questions among various subcategories.

The questions were further classified based on question types:

1) Recommendation: Questions like "Please recommend some places for food in Clemson."
2) Opinion: Questions like "What is a better programming language, PHP or Python?"
3) Factual: Questions like "How do I make my playlist private on YouTube?"
4) Rhetorical: Questions like "What is the aim of life?"

Figure 19 shows the distribution of questions based on their types. As seen from the figure, the users asked a large number of opinion-type questions. Approximately 20% of the questions were recommendation-type questions, 36% were opinion-type questions, 25% were factual-type questions, and 19% were rhetorical-type questions.

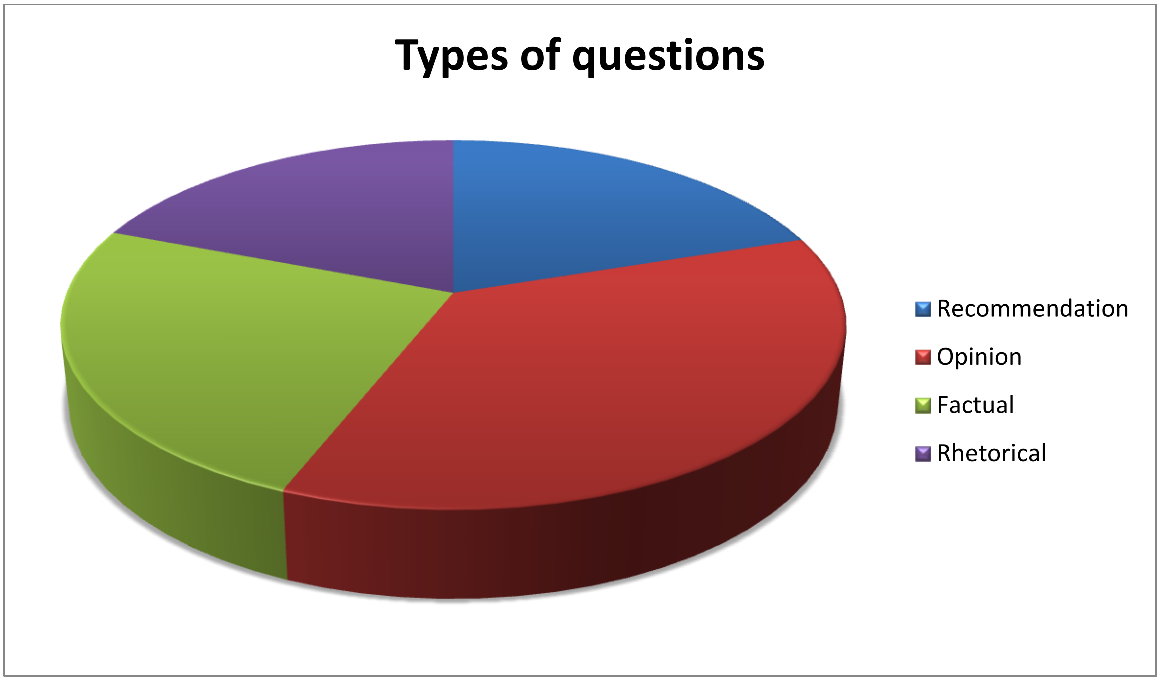


Figure 19. Distribution of questions based on their types.

Figure 20 shows the number of answers posted per question for questions with at least one response. Figure 21 shows the number of responses for questions that received at least one response. From Figure 21, it can be seen that approximately 47% of questions have just one response, and approximately 13% of questions have more than 4 responses. One observation is that most of the questions receiving only one response are factual questions, since one answer is sufficed for such questions. However, if the question asks one’s opinion, it tends to have more responses, as no answer is the final answer. For example a question like, “Should I buy a Windows laptop or MacBook?”, would have more responses than a question like ”What is the capital of Oregon?”.

***Potential benefit of SocialQ&A**: SocialQ&A provides a platform for both factual and non-factual questioning, and the opinions from social friends could be a better reference for the questioner for non-factual questions.*

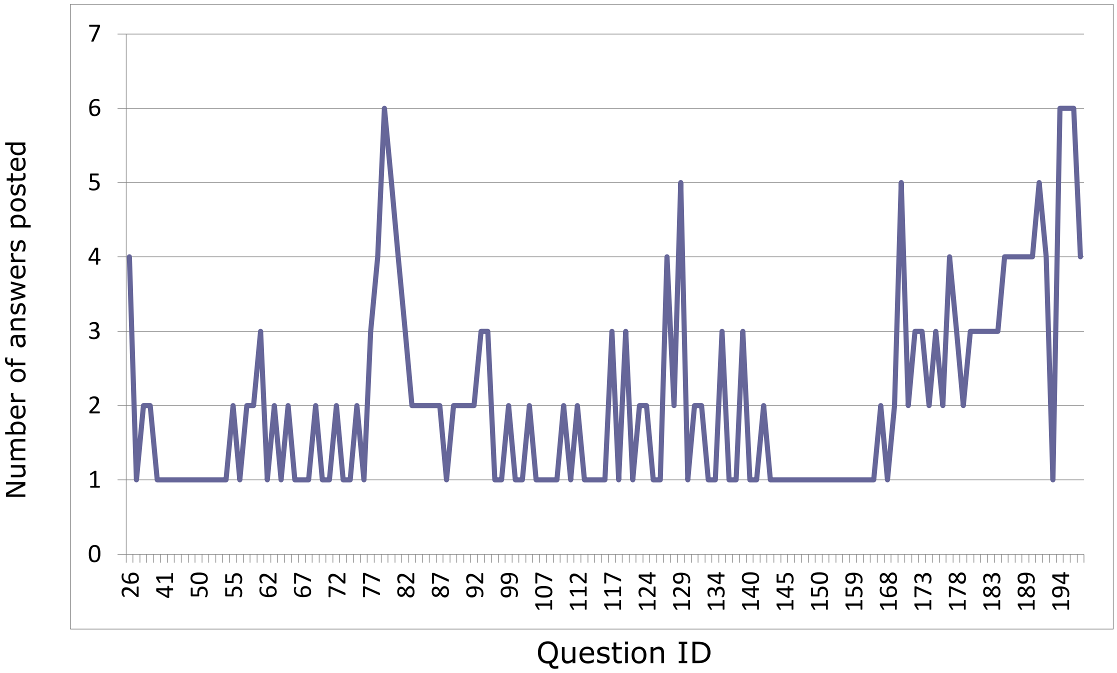


Figure 20. The number of answers received by each question.

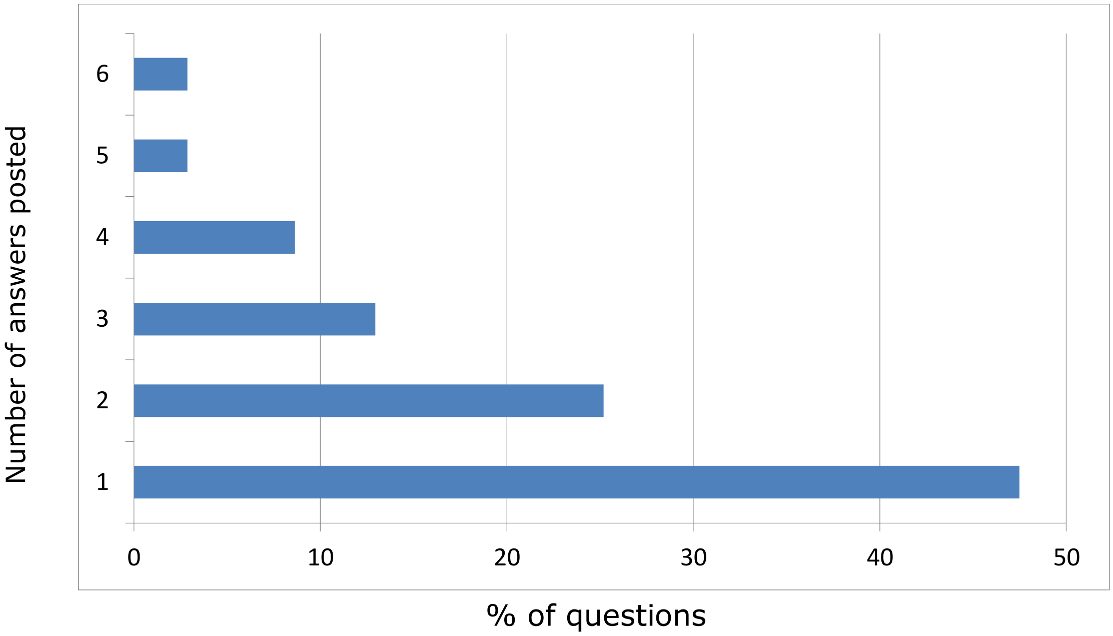


Figure 21. The number of answers received vs. % of questions.

4.3 Quality of answers

Another important metric to be considered is the quality of responses received. For every question asked, the end user was able to rate the answer on a scale of 1 to 10. The responses were stored and the following statistics were obtained. Out of 282 answers posted, the users of SocialQ&A rated 233 answers; the remaining answers remained unrated. To study the quality of answers in further detail, we calculated the average rating and the maximum rating of each question.

A single question may have multiple answers; hence, we calculated the average rating for each question and present the results in Figure 22. The results obtained from the current prototype system are promising. The average rating of all answers is 8.675, ignoring those that were not rated. The median is 9.29, the minimum is 1, and the maximum is 10.

The correlation between the question length and the question rating was also analyzed because intuitively, long questions tend to be easier to understand. Moreover, long questions help the answer provider determine what the end user is looking for, enabling him/her to provide a more accurate answer. Any question that was explained using more than one sentence is considered a long question, while the remaining questions are considered short questions. Our results show that longer questions have an average rating of 9.33, which is higher than the overall average rating.

Another way to examine the response quality is to find the maximum rating that an answer received for a particular question. The analysis of the maximum rating is meaningful because if a question received four answers, the highest rated answer provides the end user with the desired information and the other answers could be neglected. Considering this reasoning, Figure 23 plots the rating of the maximum rated answer of each question. The average maximum rating over all questions was found to be 9.05, the median was 10, the minimum was 1, and the maximum was 10.

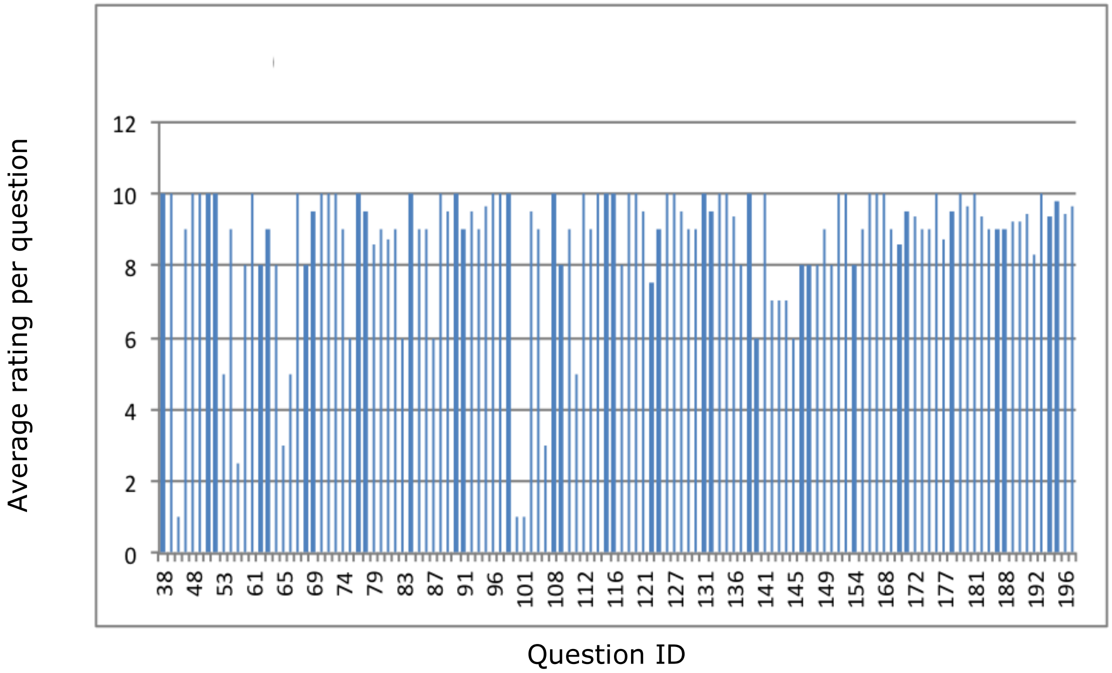


Figure 22. The average rating of each question.

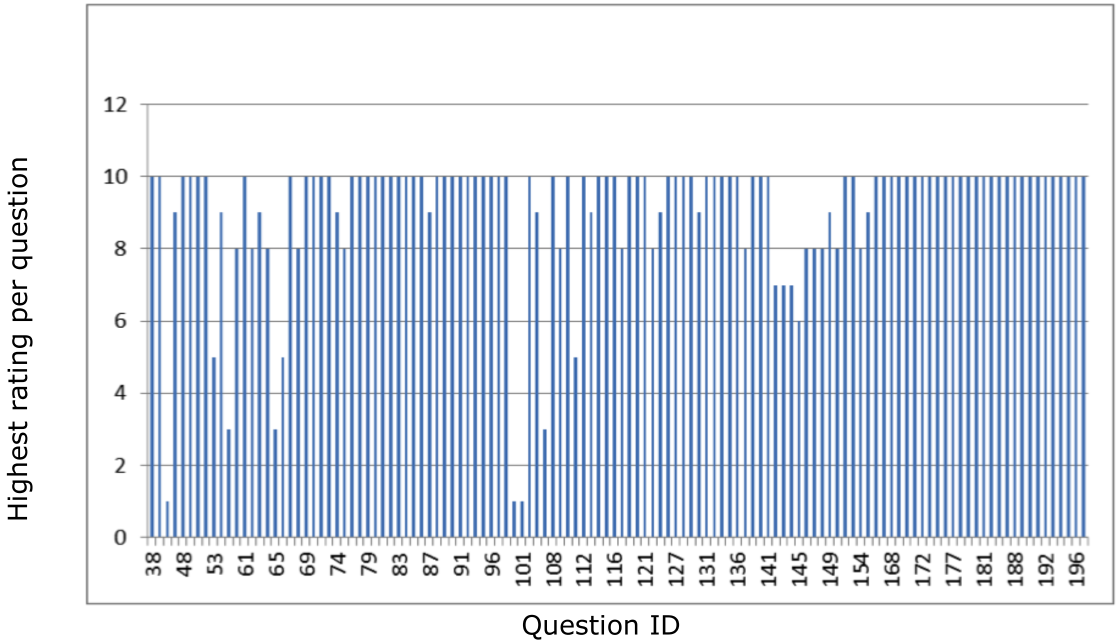


Figure 23. The maximum rating of each question.

In Yahoo! Answers, the average user rating for rated questions is around 8.44. As shown above, SocialQ&A has an average rating of 8.675, which means it performs better in terms of answer quality. Furthermore, considering the average of the highest rated answers, the average rating rises to 9.05, which is even better. It might be unfair to compare these results directly with Yahoo! Answers, since Yahoo! Answers contains hundreds of millions of users and SocialQ&A is a small system consisting of 124 users. However, the current performance of SocialQ&A is encouraging, indicating that SocialQ&A may become a better Q&A medium in the future.

The rise in ratings can be attributed to two factors: (1) since the answer provider belongs to the end user's immediate social network, he/she is highly motivated to provide better quality answers and (2) the question is mapped to the potential answer provider whose interests most closely matches the topics of the question. The result of this analysis verifies the effectiveness of our proposed algorithms: (1) User Interest Analyzer, (2) Question Categorizer, and (3) Question-User Mapper. The User Interest Analyzer algorithm can more accurately reflect the user's interests and where their posed questions belong. The Question Categorizer can more accurately derive the interest categories of questions. By mapping a question's interest categories to a users' interests, SocialQ&A can more accurately identify potential answer providers that can provide high-quality answers. In the prototype study, SocialQ&A had a very limited user set. We expect that the answer quality would be further improved as more users join SocialQ&A,

because there would be more people online at a given time and the probability that an expert exists among users also increases.

The answer quality was further analyzed based on the type of question. It was found that:

1. The avg. rating per factual question is 9.14
2. The avg. rating per opinion-type question is 8.67
3. The avg. rating per suggestion-type question is 8.18
4. The avg. rating per rhetorical-type question is 8.95

Thus, the observations indicate that factual questions have a higher average rating per question, most likely because such questions can only have one correct answer. The answer quality for rhetorical questions is determined solely by the end user's perception. Also, it can be seen that the opinion-type questions have a higher average rating than the suggestion-type questions. This is because when asking an opinion-type question, the end user typically asks for a choice between 2-4 items that he/she has shortlisted, whereas suggestion-type questions typically have a wider range of options.

4.4 Wait time for answers

Wait time is the time period between asking a question and receiving a response. Figure 24 plots the wait time for an end user to receive a response to his/her question. We see that a large percentage of questions (around

50%) are answered within 8 minutes, which is a very short amount of time. In Yahoo! Answers, less than 50% of questions receive answers within 15 minutes. As mentioned in earlier, SocialQ&A is not directly comparable to Yahoo! Answers because of the large difference in the amount of users. However, the results obtained from SocialQ&A are promising and show signs of future improvement on current Q&A systems. We also see that 15% of the questions in SocialQ&A are answered after a time period of one day for two reasons. First, due to the limited number of users in the system, sometimes the answer providers to whom the question was forwarded were not active, leaving that question unanswered until those users log in again. Second, because the number of users in the system was very small and very few users were online at a given time, some questions were left unanswered for longer periods of time.

Conversely, about 84% of the queries were answered within a day, which is a very good result for a system consisting of only 124 users. The results of the analysis again verify the effectiveness of our three proposed algorithms. By considering the social connectedness between the potential answer provider and the end user, SocialQ&A can more accurately identify potential answer providers that are willing to answer the questions within a short time. This result again suggests the promise of the SocialQ&A system, considering that the response time in a Q&A system is assumed to decrease with an increase in the number of users because the probability of a larger number of people being online at the time when a question is posted increases with the total number of users in the system. Moreover, the probability of users having expertise on a certain topic also increases with the total number of users in the system.

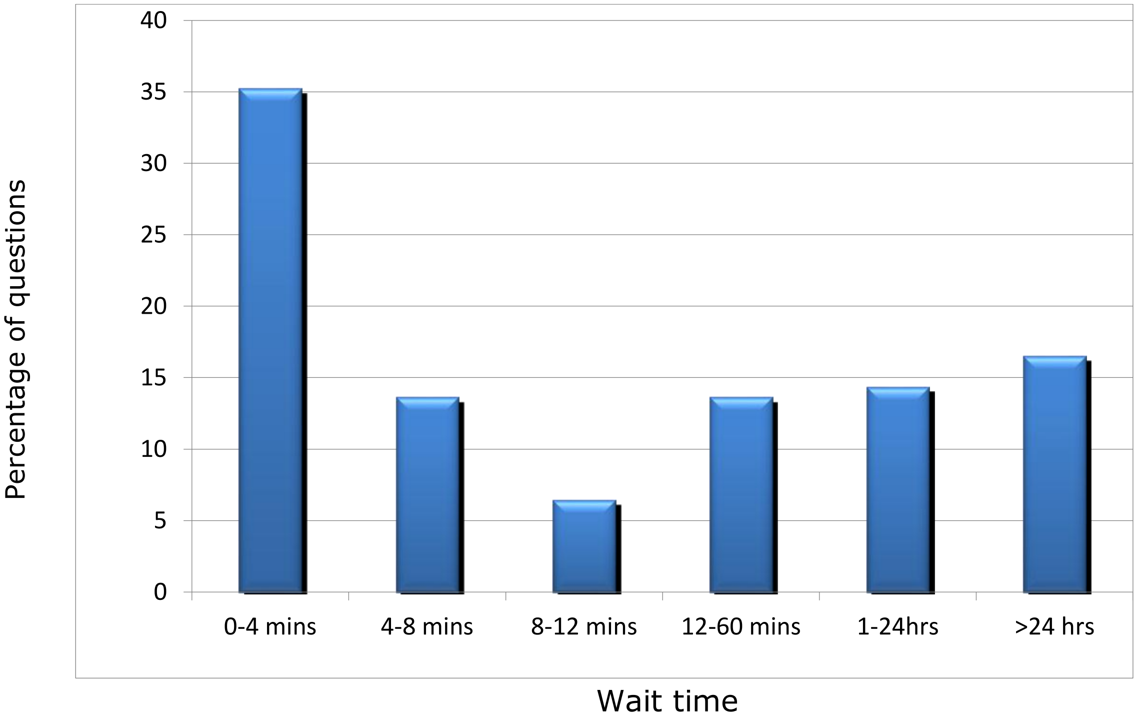


Figure 24. Percentage of resolved questions with different wait times.

The response time of answers was also analyzed based on the type of question. It was found that:

1. Most of the factual questions (around 80%) were answered within an average of 16.1mins
2. Most of the opinion-type questions (around 70%) were answered within an average of 59.87mins

3. Most of the suggestion-type questions (around 70%) were answered within an average of 71.62mins
4. Most of the rhetorical-type questions (around 70%) were answered within an average of 123.83mins

From these results, we conclude that the reason for late responses regarding the rhetorical questions is the nature of the question; conversely, factual questions get responses sooner because the answers are well established. Also, as mentioned in the previous section, the end user generally narrows down the choices for opinion-type questions; hence, they are answered faster than the closely related suggestion-type questions.

***Potential benefit of SocialQ&A**: SocialQ&A reduces the wait time of answers because as the questions are mapped to the end user's close friends who have an interest in the topics of the questions, they tend to respond quickly to the question due to the close social relationship and their expertise.*

4.5 Limitations and enhancement of SocialQ&A

We outline the limitations of SocialQ&A and possible improvements as follows.

1. The prototype test of SocialQ&A had a limited number of users in the system. Since the number of users in SocialQ&A is very small, a direct comparison between SocialQ&A and Yahoo! Answers or Baidu Zhidao

(which contain hundreds of milllions of users) might not be fair. However, the results obtained from SocialQ&A are encouraging and show that SocialQ&A could become a promising Q&A system in the future.

2. SocialQ&A has a limited number of interest categories in the system. For testing purposes, the number of major categories in the system was limited to 4 and a total of 36 categories were present in the system. In our future work, we will study the results with more categories.
3. SocialQ&A currently has a single social network rather than multiple individual social networks. However, the single social network does not affect the results because SocialQ&A only focuses on how to leverage an individual social network for better Q&A services to the users within the network. A full system with multiple individual social networks would further enhance the system performance because users from different social networks can share their historical answers stored on the server. We will implement multiple individual social networks in SocialQ&A to confirm this expectation.
4. In the current SocialQ&A system, users cannot subscribe to a particular category to receive all questions in that category. An additional feature of subscribing to a particular category could be added to further enhance the performance of SocialQ&A.

5. The current prototype of SocialQ&A does not have demographics on the users. Therefore, if all of the current users are from the same demographic, say students, this is not representative of real-world systems. In our future work, we will include demographic information to the prototype to further refine our study.

4.6 Summary

This chapter has provided the results and analysis of SocialQ&A. We have analyzed various aspects of the Q&A system, such as user activity, the number of questions and answers, quality of answers, and wait time before receiving a response to a particular question. The following chapter provides conclusions drawn from the analysis and offers some future research directions.

CHAPTER 5

CONCLUSION AND FUTURE WORK

This chapter summarizes and concludes the analysis of SocialQ&A, and provides some additional features that were identified, which could make SocialQ&A a more efficient system but have not yet been implemented. Additionally, it offers some future research directions.

5.1 Conclusion

Q&A systems are used by a large group of people for purposes such as information retrieval, academic assistance, and discussion. The growing importance of Q&A systems has led to numerous research developments that are directed toward making Q&A systems more effective. The motivation for this research is to increase the quality of answers received and decrease the wait time for answers by forwarding the questions to appropriate answer providers. Toward this goal, we have developed a social network based Q&A system, called SocialQ&A. It utilizes the strengths of a social network to forward the question to potential answer providers, ensuring that a given question receives a high-quality answer and that a given question is answered within a short period of time. Specifically, the contributions of this research can be summarized as follows:

1. We have developed the Q&A system, called SocialQ&A, which consists of three components: (1) User Interest Analyzer, (2) Question Categorizer, and (3) Question-User Mapper. These three components are new methods that enable SocialQ&A to consider user interests and user social connectedness to identify potential answer providers in order to improve the quality of answers and reduce the wait time for answers.
2. We have implemented a real-world prototype SocialQ&A system, and collected Q&A activity during one month from 124 real users in the system.
3. We have analyzed performance data obtained from the real-world prototype SocialQ&A system. Analytical results show the potential for SocialQ&A to improve on the performance of current Q&A systems.

SocialQ&A is different from previous Q&A systems in that it leverages social networks and exploits both user interests and social relationships to more accurately identify potential answer providers. Also, SocialQ&A removes the burden from answer providers by delivering the questions they might be interested in directly to them, as opposed to requiring answer providers to search through a large collection of questions to find those that he/she would be able to answer satisfactorily. SocialQ&A incorporates three novel algorithms for accurate potential answer provider identification. This research provides a promising approach to notifying the correct users in the Q&A system.

Major observations from data analysis on our small-scale prototype SocialQ&A system can be summarized as follows:

1. SocialQ&A is effective at routing questions to appropriate users by exploiting social connections and common interests; thus, it has the potential to improve the quality of answers. These three components are new methods that enable SocialQ&A to consider user interests and user social connectedness to identify potential answer providers in order to improve the quality of answers and reduce the wait time for answers.
2. SocialQ&A improves the quality and reduces the wait time of answers because as the questions are mapped to the end user's close friends, who have the interest in the topics of the questions, they tend to response quickly to the question due to the close social relationship and their expertise. A significant percentage of the questions were answered within a short amount of time (8 minutes).
3. SocialQ&A provides a platform for both factual and non-factual questioning, and the opinions from social connections may be a better reference for the questioner.

Given the amount of time the system was tested and the number of users in the system, SocialQ&A performs very well and shows a substantial improvement over existing systems. We expect that the quality of answers and the wait time in Q&A systems tend to improve with an increase in the number of users. Thus, we are optimistic that SocialQ&A has the potential to become a promising Q&A system in the future.

5.2 Future work

The algorithms implemented as a part of SocialQ&A make it a promising powerful and effective Q&A system. However, this thesis identifies some improvements that could be incorporated to make the system more usable and resourceful. This section also provides some direction for future research on Q&A systems.

One improvement that would be useful is the integration of SocialQ&A with the existing social networks like *Facebook*[8], *Twitter*[9], *Linkedin*[10], etc. Horowitz *et al.* [2] integrated this functionality into their system *Aardvark*. Such integration will empower users to utilize their existing social networks. This integration would also make tracking user interests more accurate, since it would be possible to crawl the users' statuses and posts on the social networks to dynamically update their interests. This feature will also attract more users. When an end user asks a question, the profile data of that end user's friends can be used to send question notifications to his/her friends if appropriate. Along with the question, an invitation to join the SocialQ&A platform could be sent. This integration would be a significant next step for this research.

[8] http://www.facebook.com

[9] http://www.twitter.com

[10] http://www.linkedin.com

Another idea that would be potentially advantageous is making the server-side distributed; this would allow the current system to be scalable and would increase the speed of computing the various parameters required to predict the pool of optimal potential answer providers. The Hadoop distributed file system presented in Shafer *et al.* [23] could be used to for this purpose. Hadoop has been adopted widely for the purpose of distributed data processing.

Furthermore, future research could make the system decentralized in such a way that a central server is not required and the users in the system form a peer-to-peer (P2P) structure. Decentralized search is an important research topic that would be well suited to social search, as well as searches in P2P networks as stated in Kleinberg [33] and Kleinberg and Raghavan [34]. Since most of the transactions are likely to occur among friends, the P2P networks can be modeled to exploit that feature. Condie *et al.* [36] present peer-level protocols that are adaptive and self-organizing. Likewise, Banerjee and Basu [35] have presented a social query search model that would be pertinent to this research. The intention would be to integrate SocialQ&A into a P2P system, conduct experiments with the system, and analyze the system performance as well as improve the availability of the system as a whole.

In the P2P-based SocialQ&A, if the questions and answers are stored on a client machine and if that particular client machine is unavailable, the question and answers stored on that client machine would be unavailable.

Research could be conducted to formulate an algorithm to eliminate this problem, possibly using the concept of data replication. However, the algorithm should be efficient enough to ensure that the data is available at all times with a minimum amount replication.

REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Computer Networks and ISDN Systems* 30, 1998.

[2] D. Horowitz and S. D. Kamvar. The anatomy of a large-scale social search engine, In Proceedings of the *19th international conference on World wide web,* Raleigh, North Carolina, USA, April 26-30, 2010.

[3] M. L. Radford, C. Shah, L. Mon, and R. Gazan. Stepping stones to synergy: Social Q&A and virtual reference [ASIST 2011 panel]. In Proceedings of the *American Society for Information Science and Technology*, volume 48: 1–4, 2011.

[4] Z. Gyongyi, G. Koutrika, J. Pedersen, H. Garcia-Molina. Questioning Yahoo!! Answers. *Technical Report. Stanford InfoLab*, 2007.

[5] B. Li and I. King. Routing questions to appropriate answerers in community question answering services. In Proceedings of the *19th ACM international conference on Information and knowledge management*, ACM, New York, NY, USA. 2010.

[6] B. Li, I. King, and M. R. Lyu. Question routing in community question answering: putting category in its place. In Proceedings of the 2*0th ACM international conference on Information and knowledge management (CIKM '11)*, ACM, New York, NY, USA. 2011.

[7] X. Cao, G. Cong, B. Cui, C. S. Jensen, and C. Zhang. The use of categorization information in language models for question retrieval. In Proceedings of the *18th ACM conference on Information and knowledge management (CIKM '09)*, ACM, New York, NY, USA, 2009.

[8] J. Guo, S. Xu, S. Bao, and Y. Yu. Tapping on the potential of Q&A community by recommending answer providers. In Proceedings of the *17th ACM conference on Information and knowledge management (CIKM '08)*. ACM, New York, NY, USA, 2008.

[9] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In Proceedings of the *thirteenth ACM international conference on Information and knowledge management (CIKM '04)*, ACM, New York, NY, USA, 2004.

[10] M. Liu, Y. Liu, and Q. Yang. Predicting best answerers for new questions in community question answering. In Proceedings of the *11th international conference on Web-age information management (WAIM'10),* Springer-Verlag, Berlin, Heidelberg, 127-138, 2010.

[11] M. Akiyoshi, K. Iwai, and N. Komoda. A retrieval method for similar Q&A articles of web bulletin board with relevance index derived from commercial web search engine. In Proceedings of the *10th International Conference on Information Integration and Web-based Applications & Services (iiWAS '08)*, ACM, New York, NY, USA, 2008.

[12] M. Veit and S. Herrmann. Model-View-Controller and object teams: a perfect match of paradigms. In Proceedings of *the 2nd international conference on Aspect-oriented software development (AOSD '03)*, ACM, New York, NY, USA, 2003.

[13] G. A. Miller. WordNet: a lexical database for English. *Communication. ACM* 38, 11, 39-41, November 1995.

[14] G. A. Miller. WordNet: An on-line lexical database. *International Journal of Lexicography* 3, 4, 235—312, Winter 1990.

[15] J. Kamps, Visualizing WordNet structure. In Proceedings of the *1st International Conference on Global WordNet*, 182-186, 2002.

[16] P. Han, R. Shen, and F. Yang. Intelligent Q&A System Based On Case Based Reasoning. In Proceedings of the *International Conference of Machine Learning and Cybernetics*, 345 - 348 vol.1, 2002.

[17] P. Barker. Using intranets to support teaching and learning, Innovations in *Education and Training International*, 36, 1, 3-10, 1999.

[18] R. T. Putnam and H. Borko. What Do New Views of Knowledge and Thinking Have to Say about Research on Teacher Learning? In *Educational Researcher* Vol. 29, No. 1, pp. 4-15, Jan. - Feb., 2000.

[19] G. Salomon and D. N. Perkins. Individual and social aspects of learning. *Review of Research in Education*, Vol. 23, pp. 1-24, 1998.

[20] C. Fellbaum. WordNet: an electronic lexical database, by *MIT Press*, May 1988.

[21] N. Xie and W. Liu, An Answer Fusion Model for Web-based Question Answering. In Proceedings of the *First International Conference on Semantics, Knowledge and Grid (SKG '05)*. IEEE Computer Society, Washington, DC, USA, 2005.

[22] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. *Stanford University Technical Report*, 1998.

[23] J. Shafer, S. Rixner, and A. L. Cox. The Hadoop Distributed Filesystem: Balancing Portability and Performance. In Proceedings of the *2010 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS'10)*, pages 122–133, 2010.

[24] D. Mochihashi, G. Kikui, and K. Kita. 2004. Learning non-structural distance metric by minimum cluster distortions. In *COLING-04*, 2004.

[25] D. R. Radev. A common theory of information fusion from multiple text sources step one: cross-document structure. In Proceedings of the *1st SIGdial workshop on Discourse and dialogue Vol. 10. Association for Computational Linguistics*, Stroudsburg, PA, USA, 2000.

[26] Y. Sakurai and S. Miyazaki and M. Akiyoshi. A retrieval method of similar question articles from web bulletin board. In Proceedings of the *First International Conference on Software and Data Technology*, pp.221-225, 2006.

[27] X. Liu, W. B. Croft, and M. Koll. Finding experts in community-based question-answering services. In Proceedings of the *14th ACM international conference on Information and knowledge management (CIKM '05)*. ACM, New York, NY, USA, 2005.

[28] Y. Zhou, G. Cong, B. Cui, C. S. Jensen, and J. Yao. Routing Questions to the Right Users in Online Communities. In Proceedings of the *2009 IEEE International Conference on Data Engineering (ICDE '09)*. IEEE Computer Society, Washington, DC, USA, 2009.

[29] H. Duan , Y. Cao , C. Lin , and Y. Yu. Searching questions by identifying question topic and question focus. In Proceedings of *46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2008.

[30] J. Jeon, W. B. Croft, and J. H. Lee. Finding similar questions in large question and answer archives. In Proceedings of the *14th ACM international conference on Information and knowledge management (CIKM '05)*. ACM, New York, NY, USA, 2005.

[31] K. Wang, Z. Ming, and T. Chua. A syntactic tree matching approach to finding similar questions in community-based qa services. In Proceedings of the *32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '09)*. ACM, New York, NY, USA, 187-194, 2009.

[32] M. Marchiori. The quest for correct information on the Web: hyper search engines. In *Selected papers from the sixth international conference on World Wide Web*. Elsevier Science Publishers Ltd., Essex, UK, 1225-1235, 1997.

[33] J. Kleinberg. Complex Networks and Decentralized Search Algorithms. In proceedings of the *International Congress of Mathematicians (ICM)* 2006.

[34] J. Kleinberg and P. Raghavan. Query Incentive Networks. In Proceedings of the *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS '05)*. IEEE Computer Society, Washington, DC, USA, 132-141, 2005.

[35] A. Banerjee and S. Basu. A Social Query Model for Decentralized Search. In Proceedings of *SNAKDD*, 2008.

[36] T. Condie, S. D. Kamvar, and H. Garcia-Molina. Adaptive Peer-to-Peer Topologies. In Proceedings of the *Fourth International Conference on Peer-to-Peer Computing (P2P '04)*. IEEE Computer Society, Washington, DC, USA, 53-62, 2004.

[37] B. M. Evans and E. H. Chi. Towards a Model of understanding social search. In Proceedings of the *2008 ACM conference on Computer supported cooperative work (CSCW '08)*. ACM, New York, NY, USA, 485-494, 2008.

[38] M. R. Morris, J. Teevan, and K. Panovich. What do people ask their social networks, and why?: a survey study of status message q&a behavior. In Proceedings of the *28th international conference on Human factors in computing systems (CHI '10)*. ACM, New York, NY, USA, 1739-1748, 2010.

[39] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal. Bridging the lexical chasm: statistical approaches to answer-finding. In Proceedings of the *23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '00)*. ACM, New York, NY, USA, 192-199. 2000.

[40] R. D. Burke, K. J. Hammond, V. A. Kulyukin, Steven L. Lytinen, N. Tomuro, and S. Schoenberg. Question Answering from Frequently Asked Question Files: Experiences with the FAQ Finder System. Technical Report. University of Chicago, Chicago, IL, USA, 1997.

[41] E. Sneiders. Automated Question Answering Using Question Templates That Cover the Conceptual Model of the Database. In *Proceedings of the 6th International Conference on Applications of Natural Language to Information Systems-Revised Papers* (NLDB '02), Springer-Verlag, London, UK, UK, 235-239, 2002.

[42] X Quan, L. Wenyin, and B. Qiu. Term Weighting Schemes for Question Categorization. In *IEEE Transaction Pattern Analysis and Machine Intelligence* 33, 5, 1009-1021. May 2011.

[43] W. Song, L. Wenyin, N. Gu, X. Quan, and T. Hao. Automatic categorization of questions for user-interactive question answering. *Information Processing & Management*, Volume 47, Issue 2, Pages 147–156, March 2011.

[44] S. K. Ray, S. Singh, and B. P. Joshi. A semantic approach for question classification using WordNet and Wikipedia. *Pattern Recognition Letters* 31, 13 (October 2010), 1935-1943, 2010.

[45] L. Hong, Z. Yang, and B. D. Davison. Incorporating Participant Reputation in Community-Driven Question Answering Systems. In Proceedings of the *2009 International Conference on Computational Science and Engineering* - Volume 04 (CSE '09), Vol. 4. IEEE Computer Society, Washington, DC, USA, 475-480, 2009.

[46] W. Chen, Q. Zeng, W. Liu, and T. Hao. A user reputation Model for a user-interactive question answering system. *Research Articles. Concurrent Computing: Practice and Experience*. 19, 15, 2091-2103, October 2007.

[47] Y. R. Tausczik and J. W. Pennebaker. Predicting the perceived quality of online mathematics contributions from users' reputations. In Proceedings of the *2011 annual conference on Human factors in computing systems (CHI '11)*. ACM, New York, NY, USA, 1885-1888, 2011.

[48] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993-1022, March 2003.

[49] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. ACM Trans. Inf. Syst. 22, 2 (April 2004), 179-214, 2004.

[50] J. Hwang, S. Lay, and A. Lippman. Nonparametric Multivariate Density Estimation: A Comparative Study, In *IEEE Transactions On Signal Processing*, Vol. 42, No. 10, 2795 - 2810, October 1994.

[51] G. E. Krasner and S. T. Pope. A cookbook for using the Model-View Controller user interface paradigm in *Smalltalk-80. J. Object Oriented Program*. 1, 3, 26-49, August 1988.

[52] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In Proceedings of the *international conference on Web search and web data mining* (WSDM '08). ACM, New York, NY, USA, 183-194, 2008.

[53] J. Jeon, W. B. Croft, J. H. Lee, and S. Park. A framework to predict the quality of answers with non-textual features. In Proceedings of the *29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '06)*. ACM, New York, NY, USA, 228-235, 2006.

[54] M. Surdeanu, M. Ciaramita and H. Zaragoza. Learning to rank answers on large online QA collections. In *ACL*, 2008.

[55] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In Proceedings of the *17th international conference on World Wide Web (WWW '08)*. ACM, New York, NY, USA, 665-674, 2008.

[56] V. Jijkoun and M. de Rijke. Retrieving answers from frequently asked questions pages on the web. In Proceedings of the *14th ACM international conference on Information and knowledge management (CIKM '05)*. ACM, New York, NY, USA, 76-83, 2005.

[57] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In Proceedings of the *16th international conference on World Wide Web (WWW '07)*. ACM, New York, NY, USA, 221-230, 2007.

[58] J. Bian, Y. Liu, D. Zhou, E. Agichtein, and H. Zha. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In Proceedings of the *18th international conference on World wide web (WWW '09)*. ACM, New York, NY, USA, 51-60, 2009.

[59] P. Jurczyk and E. Agichtein. Discovering authorities in question answer communities by using link analysis. In Proceedings of the *sixteenth ACM conference on Conference on information and knowledge management (CIKM '07)*. ACM, New York, NY, USA, 919-922, 2007.

[60] M. Bouguessa, B. Dumoulin, and S. Wang. Identifying authoritative actors in question-answering forums: the case of Yahoo! answers. In Proceedings of the *14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08)*. ACM, New York, NY, USA, 866-874, 2008.

[61] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In Proceedings of the *17th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '94). New York, NY, USA, 232-241.

[62] E. Pennisi. How Did Cooperative Behavior Evolve? In *Science (July '05). Vol. 309 no.5731.* 93, 2005.

[63] Computer Science vs Computer Engineering, http://www.eng.buffalo.edu/undergrad/academics/degrees/cs-vs-cen [accessed in June, 2012].

[64] Z. Li, H. Shen, G. Liu and J. Li. SOS: A Distributed Context-Aware Question Answering System Based on Social Networks, *Proc. of IEEE International Conference on Distributed Computing Systems (ICDCS)*, June 18-21, Macau, China, 2012.