5-2015

Exploring The Effects Of Multimedia Content On A Question And Answer System

Adam Ross Carroll
Clemson University

Exploring the Effects of Multimedia Content on a Question and Answer System

---

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
Adam Ross Carroll
May 2015

Accepted by:
Dr. Haiying (Helen) Shen, Committee Chair
Dr. Robert Schalkoff
Dr. Brian Malloy

Abstract

Online Question and Answer (Q&A) websites have been a part of the Internet for many years, the most well-known being Yahoo! Answers. These websites allow us to ask and answer questions with other Internet users around the world, utilizing one of the Internet's greatest strengths: information sharing. In the past, this sharing was restricted to primarily text-based content, and previous research has been almost solely devoted to these text-based Q&A systems. This research includes topics such as improving text-based question quality, and methods for finding the best text-based answer. Now, new Q&A systems such as Jelly have recently been released that utilize multimedia, including images, audio, or video for questions and/or answers. These new multimedia features were not part of previous text-based research, and the objective of this project is to fill that research void: how does the use of multimedia in a question affect the odds of receiving the correct answer? Our findings will not only affect this previous Q&A research on question quality and forwarding, but also reveal new security and privacy concerns. To perform our research, we created several different types of multimedia questions, spanning many different topics. We then studied how users answered each question, using our custom MultiQuery website to facilitate the experiment. This website is unique to Q&A in that it allows for many different multimedia question types, not simply text. Once the experiment was completed, we analyzed the results and determined if and to what degree multimedia use in a question improves or worsens the quality of answers; both overall and for each specific topic of questions. We found that multimedia did in fact have a beneficial effect, especially with images, showing a higher answer rating and correctness percentage for most multimedia types. We also

found many security and privacy risks from integrating multimedia into a Q&A website, including loss of privacy through image sharing and voice recognition, as well as security dangers to Q&A system owners through illegal, malicious, or explicit uploaded multimedia content. Our analysis and discussion will help future multimedia Q&A systems as they seek to implement the most effective forms of multimedia, will aid future research into multimedia question quality and forwarding algorithms, and will provide recommendations for safer Q&A security practices that account for multimedia.

Dedication

I would like to dedicate this thesis to my closest friends and family: my father Darren Carroll, who exampled a tireless work ethic and taught me how to do my best; my mother April Carroll, who homeschooled me all the way through high school and has helped me by proofreading this and many other papers; my brother Austin Carroll, who has helped me technologically with many projects such as this one; and my fiancee Kara Felber, for encouraging me not to quit and being my cheerleader every day. Finally, and most importantly, I would like to thank my Lord and Savior Jesus Christ who has blessed me in countless ways, including my schoolwork, my family, and my future wife.

Acknowledgments

Many professors, classmates, friends, and family members were involved in this project, and I would like to thank all of them for their support and hard work that made this thesis possible.

I would like to thank my adviser and committee chair Dr. Haiying Shen for her help throughout the entire project. I would also like to thank my committee members Dr. Robert Schalkoff and Dr. Brian Malloy for their investment in me both in the classroom, and in the time spent reading, analyzing, and helping me improve this thesis. Thank you to Guoxin Liu as well, for his help throughout multiple rounds of revision.

I would like to thank all of the participants in this study; thank you for your hard work every week. You were always punctual in your assignments, never missing a week, and you showed a true desire to make this project a success; I could not have asked for more diligent participants.

Finally, but definitely not least, I would like to thank Josh Johnston for helping me create the MultiQuery website and for helping me write questions every week. You and I spent an entire, busy semester simply building the website. Without your user interface skills and programming help, I would never have been able to even begin this project.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter, we introduce the research and provide the motivation for the project. We present the background and history of Question and Answer (Q&A) systems, their recent technological and multimedia trends, security and privacy risks stemming from these trends, and finally the objectives of and contributions provided by our research.

1.1 Background of Q&A Systems

With the advent of the modern Internet in the 1990's [82], information sharing became the easiest that it had ever been. Suddenly, users across the world could communicate information in mere seconds. But as Internet usage and the amount of available content grew, it became increasingly difficult to find specific content amongst all of the other available information. To solve this problem, search engines such as Google [12] and Bing [8] were developed that take specific keywords supplied by users and scour the Internet for websites and other content matching those keywords. These search engines are excellent for finding answers to questions with objective, factual answers, but what about subjective, non-factual questions? Or, what if your objective question has not yet been completely answered online?

This problem was solved by the development of Q&A systems such as Yahoo! Answers [26]. These websites allow users to ask questions to and answer questions from other Internet users around the world. Yahoo! Answers allows for a wide range of question categories, but other Q&A websites are for much more specified topics. StackOverflow [19] is for questions in specific categories such as computer programming, Jobstr [15] for questions about jobs, etc. This gives Internet users many options in their search for answers.

These websites worked well for past generations when the early Internet operated almost solely as a text-based medium, but broadband Internet and mobile technologies are beginning to change our habits and therefore the way we ask and answer questions online. As smartphone ownership increases, more and more users are always connected to the Internet and able to use Q&A systems. According to Pew Research Center, smartphone ownership rates have risen from 35% to 58% in just the past 3 years [75] (see Figure 1.1); that is a 66% increase! These new technologies enable the faster transmission of high-bandwidth images, audio, and video, which are quickly becoming a part of Q&A. According to Merriam Webster, multimedia is defined as, "a technique (as the combining of sound, video, and text) for expressing ideas (as in communication, entertainment, or art) in which several media are employed" [35]. Essentially, multimedia is combinations of the aforementioned content types. Each form of multimedia content can be classified as a multimedia "type", such as image or audio.

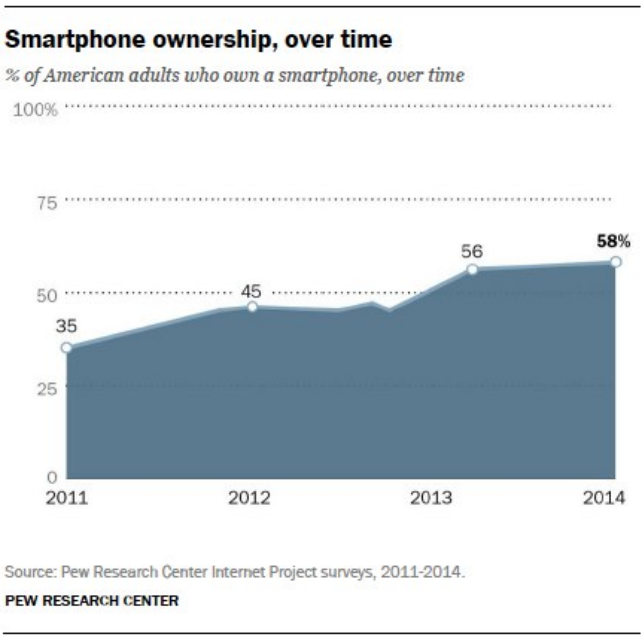

Figure 1.1: American smartphone ownership over the past three years [75]

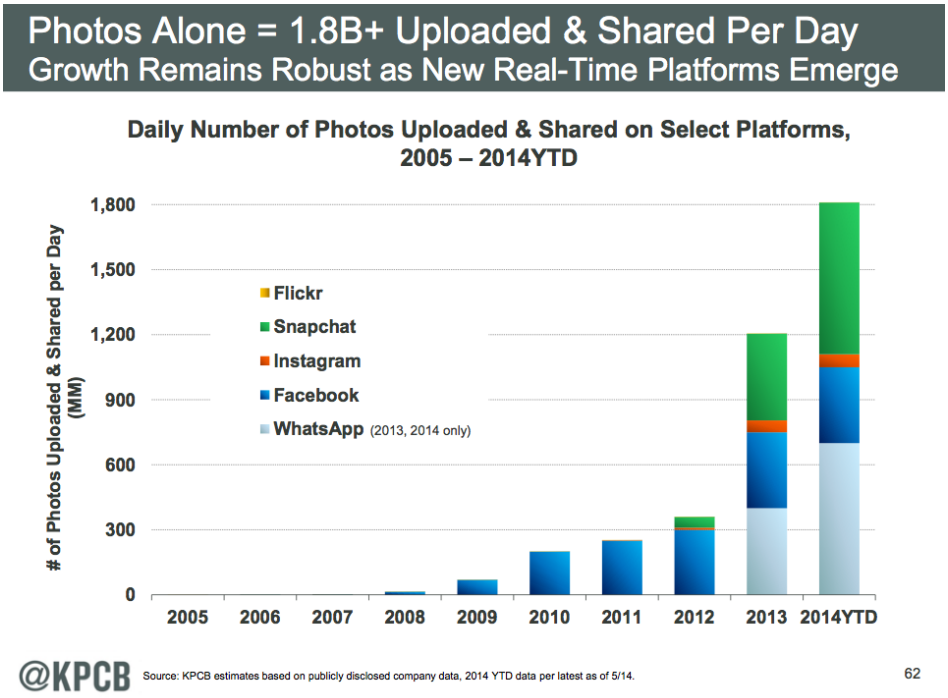

Figure 1.2: Photos uploaded and shared per day over the past ten years [64]

This use of multimedia is becoming a popular means of communication. As reported by venture capital firm KPCB in Figure 1.2, an estimated 1.8 billion photos were uploaded and shared per day in 2014 [64], and that is only from examining 5 social network and image sharing platforms! Modern Q&A systems, not to be outdone by their social network counterparts, are integrating multimedia technologies to provide novelty and entice users. These Q&A systems will provide an outlet for the image uploads into a Q&A format.

These modern Q&A systems include mobile applications (apps) such as Jelly [14] and Ask.fm [6], which allow users to ask and/or answer questions using images and videos respectively. This opens the door for many new types of questions, such as asking other users to identify items in an image, etc. As more and more users are acquiring a smartphone with almost unlimited Internet access, usage of these apps and websites will only continue to grow. Many of these websites and apps that we have mentioned incorporate social networking features, which is a common trend for Q&A systems. Because of this commonality, our method of study also uses a social network system.

Additionally, as Q&A systems evolve, new security risks are beginning to appear. With the advent of camera phones and geolocation data stored within the captured images, it is now possible for malicious users to track anyone sharing images online that contain this data [78]. Geolocation data, combined with new (and even old) facial and audio recognition technologies, makes any image or audio file uploaded for Q&A a security risk to the uploader [47]. There are also risks for the Q&A website owners (webmasters) who must consider security, legal, and privacy concerns stemming from the uploaded multimedia content [74]. In addition to our research regarding multimedia questions, this thesis will also examine the risks and consequences involved with uploading and permitting the upload of these images and audio files to a Q&A website, as well as steps that can be taken to mitigate these risks for both the users and webmasters.

1.2 Our Research

Research in the realm of Q&A networks has been fairly substantial. There have been many papers published on improving question quality [49, 54, 58, 79, 84], methods for forwarding questions to users best able to answer them [46, 61], finding "expert" users [41, 87], etc., which will all be discussed more thoroughly in the following chapter. The problem is that most of this research is focused on traditional, text-based Q&A systems, and does not account for the effect of multimedia on questions and answers. There is a need for further research into the implications of multimedia on questions and answers, in terms of both their practicality and safety.

1.3 Thesis Objective

In this thesis, the term "asker" represents the user who is asking the question, "answerer" represents the user who is answering the question, and "user" generically represents both. Our objective is to answer the following research questions:

1. How does the use of images or audio (multimedia) in a question affect the correctness of answers? Studying this will show whether or not integrating multimedia content is worth the added work and risk.

2. Between images and audio, which of these two multimedia types is more effective in receiving correct answers? This will help determine that, if multimedia is going to be used, what should be the preferred or first choice.

3. When answering a question, what forms of multimedia (if any) do answerers prefer the asker to be using, and to what degree is this dependent on the specific answerer? Studying this will show how likely users are to answer questions of each multimedia type, and whether or not this is true for all answerers.

4. What categories of question (if any) are the most suited to a multimedia Q&A website? Because multimedia Q&A is new, we do not know which categories of questions will be most suited for the use of multimedia. This will help specialized Q&A systems, such as StackOverflow, in determining if their domain should integrate multimedia.

5. What are the security and privacy risks involved with a multimedia Q&A website? Security and privacy should be studied beforehand, not after multimedia is already implemented. Answering this question will help Q&A websites avoid these previously unconsidered risks.

Of these research questions, comparing #2 and #3 is very important. While an asker may receive better answers from an audio-only question, the answerer may loath it because of the added difficulty involved in listening to an entire audio clip. This may cause the answerer to avoid these types of questions altogether, making it harder for the asker to find answers with that multimedia type. These questions will help our research team as we expand on text-only question forwarding and rating algorithms, as well as commercial Q&A networks that are seeking the best multimedia choices for improving their website

or app. We provide answers to these questions through the use of a custom Q&A system named MultiQuery that we designed and implemented. The details of this website will be described in Chapter 3.

1.4 Thesis Contributions

The contributions of this thesis are as follows:

1. The design and implementation of a multimedia Q&A System. MultiQuery is a website designed specifically for our multimedia Q&A experiment. It allows users to ask and answer questions, rate questions, follow other users, etc. Built using the Ruby on Rails framework [18], this is a very maintainable and expandable website. And, while many Q&A systems such as Jelly and Ask.fm allow for image or audio questions exclusively, we allow for the posting of text, image, and audio questions all on the same system.

2. The analysis of the experiment's data. We have analyzed the average answer rating and correctness percentage for each multimedia type and question category, both overall and for each specific user, the users' preferred and perceived best multimedia type to answer, what topics of questions are best suited to each multimedia type, how well users answered multimedia questions in their chosen category of expertise, and how users respond to misleading questions.

3. A discussion and analysis of multimedia security and privacy risks. Adding multimedia functionality to a Q&A website creates new risks that must be understood and addressed. These include privacy risks to users from images and audio, as well as security risks to website owners from illegal, malicious, and explicit user uploads.

1.5 Summary

In this chapter, we have provided the background of Q&A websites, shown their trends toward multimedia content and lack of research in this area, presented our research questions, introduced our Q&A system, and listed our research contributions. The remainder of this thesis is organized as follows. Chapter 2 provides an overview of current commercial Q&A systems and related work. Chapter 3 presents the design of our Q&A system, the experiment that was conducted, and the security study that was performed. Chapter 4 discusses and analyzes the results of the experiment through the use of charts and tables, provides findings from the user surveys, and examines the findings from our security study. Finally, Chapter 5 concludes the thesis and provides our avenues for future work.

Chapter 2

Background and Related Work

The purpose of this chapter is to familiarize the reader with features of both traditional and modern Q&A systems, provide an overview of potential multimedia security risks, and detail previous research that has been performed on these systems. We will begin by looking at traditional Q&A websites such as Yahoo! Answers and StackOverflow, and then move to modern mobile apps such as Jelly and Ask.fm, before previewing the various potential security and privacy risks from the uploaded multimedia content. Finally, we will dive into previous research on Q&A systems in general, question quality, question forwarding algorithms, classification of multimedia content, and privacy risks.

2.1 Modern Q&A Systems

Since its founding in 2005 [37], Yahoo! Answers has been the standard for a Q&A website. Users are able to ask questions to other users, and these questions are assigned a category that makes them easier to locate and browse. As users answer questions, they acquire points and higher "levels" based on those points. An asker is able to earn and award points to an answerer by marking an answer as the "best", which is another statistic tracked on a user's profile. Figure 2.1 shows an example profile, and how the statistics are utilized. These statistics and sense of reward are a major driver toward high user participation.

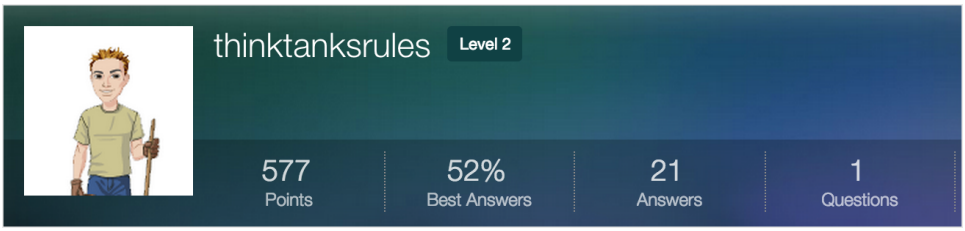

Figure 2.1: An example Yahoo! Answers user profile

There are many competitors to Yahoo! Answers, a notable one in the computer engineering and programming fields being StackOverflow [19]. Owned by a parent company hosting dozens of Q&A websites tailored to specific fields, StackOverflow specializes in questions specifically related to computer programming. The advantage here is that all of their users should be somewhat knowledgeable about programming, which results in more specific questions, and therefore better overall answers. It also is more appealing to field experts, who would rather invest their time in answering more specific questions.

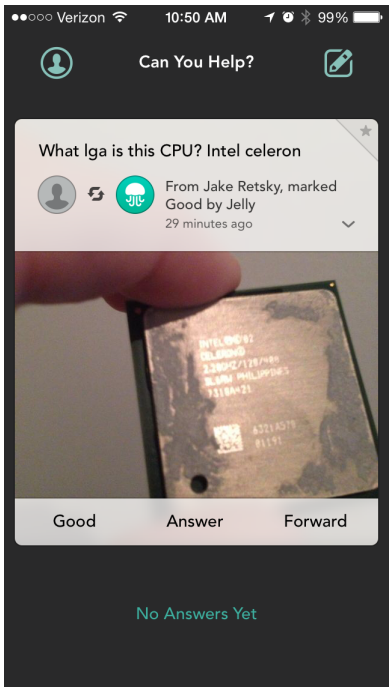

Figure 2.2: An example question on the Jelly app

These traditional Q&A systems work well, but they are all text-based. This limits the questions to containing only typed words; images and video are only possible if a hyperlink is used. But this hyperlink is inherently unsafe, as the user must navigate to another, possibly malicious website to view the additional content. In order to solve this problem, some new Q&A systems have been released, one of the most well-known being Jelly. This mobile app allows askers to upload an image along with their text-based question. The image can be of great aid in answering the question, especially questions where the asker is attempting to identify an object. Potential answerers are able to

browse questions in categories of their choice, and either answer the question, forward the question to a knowledgeable friend, or simply mark the question as "good" or otherwise well written. The more good marks that a question receives, the more likely it will be presented to a potential answerer. And all of this is implemented into a fun, mobile interface. See Figure 2.2 for an example of a question on Jelly.

Another new system is Ask.fm, a mobile app that allows a user to ask another user a text-based question and receive an image or video answer. They integrate both types of multimedia into a Q&A system, and also add social networking features to enhance user interaction. These systems are very useful and growing in popularity, but neither fully integrate multimedia as part of the question itself. Jelly does allow for images in the question, but no audio, while Ask.fm has no multimedia in the question whatsoever. These systems only scratch the surface of what is possible with multimedia in Q&A, and we seek to expand on their progress by providing a system that allows for questions with either images or audio, all contained within the same website.

2.2 Risks of Q&A

Q&A social networks have great potential, allowing almost any Internet user across the globe to ask and answer questions with other users using a standardized platform. While this seems like a fairly innocent activity, pairing these questions and answers with other (potentially leaked) information about the (assumed) anonymous users can lead to their identification, and therefore a violation of their privacy [56]. One of the greatest features of the early Internet was its anonymity, and this threat of identification is a direct attack on that assumption.

With more Internet access and increasingly faster Internet speeds becoming available, the multimedia Q&A trend is continuing to grow [62]. This introduces legal, security, and privacy concerns if the data is copyrighted, secret, or if it allows for the identification of the asker or other users. Use of these Q&A systems is being conducted more and more on

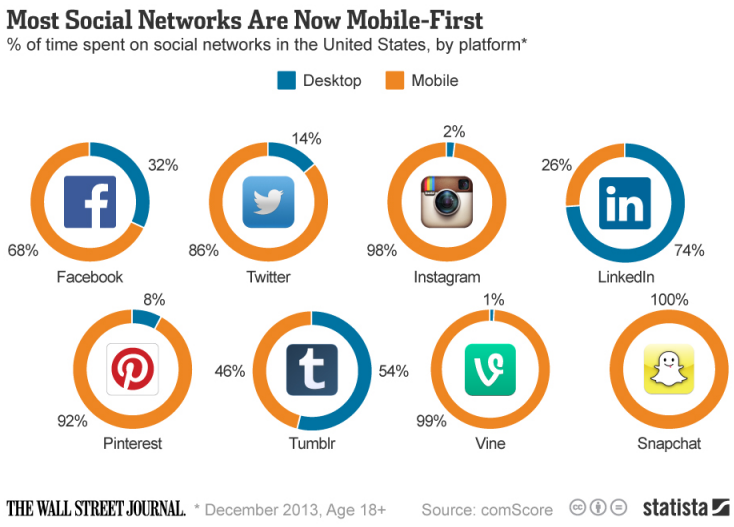

Figure 2.3: Percentage of social network use via mobile [63]

mobile phones (see Figure 2.3), and users often do not realize the additional security risks involved with these devices. When most smartphones capture an image via the phone's camera, that image is tagged with several items of "metadata", including the GPS coordinates where the image was taken and several other key attributes of the phone itself [77]. By posting image questions that contain this metadata, askers expose their location and information about their phone or camera. On top of this, great strides have been made in facial recognition technology. Facebook, for instance, has sophisticated algorithms that are able to detect faces in a image and match them to other users with surprising accuracy [71].

In addition, audio recordings have their own risks, and these are not solely restricted to mobile phones. One of the greatest risks here is that of voice recognition. When an asker uploads a video or audio file containing their voice, technology has advanced (and is continuing to advance) that will allow for the identification of the person who is speaking in the recording [21]. This, combined with the risks from shared images, should make all website users think twice before uploading anything that may compromise their privacy. The following chapters will provide a more thorough examination of each of these risks, as well as ways to prevent them.

Finally, Q&A webmasters must consider the legal and security risks of this posted content. When askers are given the ability to upload any image or audio file to a website's servers that they wish, this opens the door for questionable activity. The uploaded multimedia files may be illegal (copyrighted music, images, etc.), malicious (viruses, online scams, etc.), or explicit (pornography, sexual content, etc.) [66, 67]. Appropriate safeguards must be implemented to mitigate these risks and resolve them if they do occur.

2.3 Previous Research and Related Work

Because the integration of multimedia into Q&A systems is very new, there has yet to be much research invested in the topic; that is the motivation for this study. However, there has been much research performed on traditional Q&A systems, of particular interest to us is that involving question quality. Besides quality in general, this also involves question forwarding and classification of content. In this section, we will look at those relevant works, as well as what others have published in regards to the privacy concerns of Q&A multimedia.

Shah *et al.* [79] presented a thorough summary on Q&A systems as a whole. They mentioned three types of Q&A services: (1) "digital reference services" where the user asks questions directly to a reference librarian, (2) "expert services" where questions are answered by qualified individuals in that field, and (3) "social Q&A" where everything is community-based and answers are sought from anyone. Q&A can have different definitions depending on context, sometimes asking the question to thousands of other users on a social network, but other times simply using an online search engine without the interaction of any other users at all. A relevant research question they pose is, "What are the factors that need to be considered when measuring the quality of questions and answers?" We suggest that multimedia is one of those factors, and will later examine what aspects of multimedia are the most significant.

Action	Points
Begin participating on Yahoo Answers	One Time: 100
Ask a question	-5
Choose a best answer for your question	3
Answer a question	2
Self-deleting an answer	-2
Log in to Yahoo Answers	Once daily: 1
Have your answer selected as the best answer	10
Receive a "thumbs-up" rating on a best answer that you wrote (up to 50 thumbs-up are counted)	1 per "thumbs-up"
Receive a violation	-10

Figure 2.4: Yahoo! Answers' current point system

Yahoo! Answers is one of the most frequently cited Q&A systems, and Shen *et al.* [80] provided an in-depth look at the website's functionality. They also performed a thorough study of the interactions between users, showing how they behaved and what influenced this behavior. For instance, users who acquired a higher number of points typically did not ask many questions because of the small point cost Yahoo! Answers charged when asking a question. See Figure 2.4 for a simplified overview of Yahoo! Answers' current point system. This behavioral study could be expanded to include how multimedia affects behavior, such as what types of multimedia these members would typically use, etc. Also, are there methods in which points can be awarded or charged for using multimedia in a question? Either awarding because of the extra effort involved, or charging because of the bandwidth costs? This decision about points implementation will largely depend on how (un)important this study shows multimedia content to be in receiving the correct answer to a question.

There has been much research invested into question quality for traditional, text-based questions. Li *et al.* [58] emphasized the importance of this, explaining that not only will a low question quality affect the answer quality, but higher quality questions will also help the Q&A service grow and therefore help others to find questions relevant to their own problems. Both Harper *et al.* [49] and Jeon *et al.* [54] emphasized price as one of the greatest factors in question quality. Websites that charged for access had more involved and

committed users, and this typically resulted in better answers. Another important factor that Harper *et al.* cited is the need for a larger community; the more people that could answer a question, the better the chances of receiving the correct answer. Teevan *et al.* [84] went even further through several observations, all with the premise that the best questions were those that were well phrased. This means stating the information in the form of a question (instead of a statement), addressing the audience specifically, and using only one sentence. These three observations, particularly the last, appear very supportive of the use of images in or as a question due to its conciseness.

Another related research topic is question forwarding. When a question is asked, the goal is to present it to those that have the greatest ability to answer it (experts), and to do so in a timely manner. An expert is a user that is skilled in answering questions in specific categories, therefore the final target of question forwarding. This can either be accomplished manually by a user forwarding the question to those that they themselves know to be experts, or automatically by an algorithm that ranks users according to their expertise. The algorithmic identification of these experts is a difficult task, and many researchers have presented methods for solving it. Yang *et al.* [87] proposed their Topic Expertise Model (TEM) and Community Question Answering Rank (CQARank) to "measure user interests and expertise score under different topics." These tools can be used to find "experts with both similar topical preference and high topical expertise." Similarly, Bouguessa *et al.* [41] proposed a systematic approach to find "authoritative" or expert users through recording the number of best answers that a user supplied, and using this information to generate their "authority score". Finally, Zhang *et al.* [88] demonstrated a mechanism that automatically found suitable experts through a chain of acquaintances, as opposed to searching an entire database of users. Identifying expert users is vital to a successful question forwarding algorithm, as those users are the most likely to supply a good answer. But it is also a challenge, becoming even more difficult as new features are added to Q&A systems. Multimedia is one of those new features, and future research will need to take into account a user's abilities to answer questions of each multimedia type. Our study hopes to

show which multimedia types are the most significant, and should therefore be independent parts of expert identification algorithms.

Once experts have been determined overall, the next area of research is how to get these questions to the best experts for a specific question. Li *et al.* [61] has examined this through the eye of a distributed system. They described a lightweight system that was able to efficiently process these forwarding algorithms without resorting to the server, reducing search and computational costs. They also described their methods of string parsing using first-order logic that could match a question to possible answerers among the asker's list of friends. An additional option, as Carretero *et al.* [44] demonstrated, is a geographical-based model that looks for experts in nearby areas. This could be very useful for audio questions on a global Q&A website, where users must be found that speak the same language or dialect. Li *et al.* [60] provided a similar tool, FindU, for finding other website users in a nearby geographic area, but also preserving user privacy by keeping sensitive information private.

Diaconita *et al.* [46] presented an applied study in question forwarding, and were able to successfully implement a context-aware system. Their algorithms examined current user activity, with the intent that questions should only be forwarded to those who are currently available to answer them at that time. This concept should be a vital part of any question forwarding algorithm, especially for multimedia questions where the asker may need an immediate response to identify an object that they are considering purchasing, etc. Question forwarding algorithms will need to be enhanced to account for the addition of multimedia, and our study will help in identifying what aspects of multimedia questions should be analyzed by these algorithms.

An even simpler solution than forwarding questions to a user for answering is automatically answering the question with answers from previous, similar questions. The TREC-8 (Text REtrieval Conference) Question Answering track in 1999 was one of the first large-scale attempts at this, seeking to answer fact-based, short-answer questions [85]. This TREC program has undergone constant development, which continues to this day [23].

Some questions are naturally more suited to this type of system, especially those that are fact-based. Bian *et al.* [38] presented an effective method for answering these fact-based questions, using Yahoo! Answers as their test platform. To answer a new question, they searched the website's archives of past questions and answers, recorded the number of "plus" or "minus" votes that a past answer had received, translated these results into equations, and used this data to create a self-learning ranking function that provided the most likely answer. Suryanto *et al.* [83] went further and suggested that both an answerer's expertise and past answer quality should also be considered instead of simply matching based on relevance criteria. They provided and examined several methods for doing this, as did Bian *et al.* [39]. Many of these systems attempt to combine multiple past answers into a new answer, but Shtok *et al.* [81] attempted to simply reuse past answers "as is". When their search algorithms reached a desired confidence level, the "new" answer was created by simply copying the "old" answer. They found that their posted answers outperformed the average asker satisfaction level received by human answerers, demonstrating that past answers do not necessarily need to be synthesized and combined with others to effectively answer a question.

In terms of multimedia questions, Hong *et al.* [53] presented a preliminary study on how best to apply multimedia content as a direct answer to a question, especially through search engines. Modern search engines are optimized for returning queries of text-based media, but these search engines also require methods for returning suitable multimedia content. While some multimedia may have tags or other associated textual information that can be searched, this information is often not present or is even incorrect. This means that search engines for Multimedia Q&A (MMQA) need algorithms to match the query or question to the correct multimedia content. Once this is done, the multimedia and textual results need to be combined and ranked against each other to determine the top search results. Yadav *et al.* [86] discussed methods of data mining that found unstructured text, patterns, or relationships between multimedia content which could then be supplied as search results. Similarly, Mei *et al.* [65] presented a survey of re-ranking systems that performed multi-

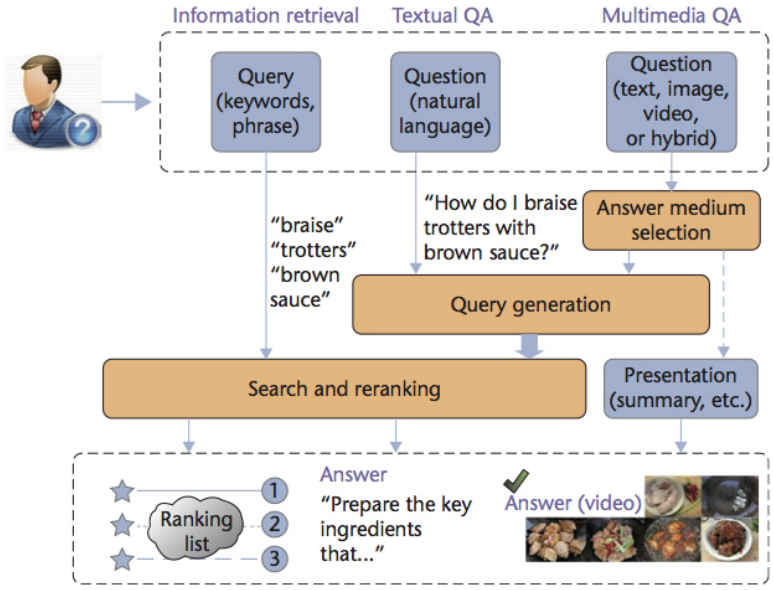

Figure 2.5: Steps for integrating multimedia content into search results [53]

media search through combinations of queries or some form of post-processing. These tools are commonly used by Google, Bing, and other search engines in image-based web searches.

Once the search results are returned, Figure 2.5 shows the steps that Hong *et al.* [53] suggested for the process of integrating multimedia into search results. They first identified the best type of multimedia for answering the question (if any), found text-based answers to combine with the multimedia, and finally ranked and presented these results. Li *et al.* [59] discussed this as it applies to video content, using their Video Reference tool to answer "how-to" Yahoo! Answers questions by posting videos directly from YouTube as the answer. Nie *et al.* [70] used their own tools to examine and parse a user supplied answer to determine if the addition of multimedia content would be helpful, and if so, added the aforementioned multimedia to the existing answer. Finally, Abu-Naim *et al.* [36] examined methods for automatically identifying multimedia that shared a common "Subject of Interest" with other users, and converted it into a format suitable for distribution. This would be useful not only for multimedia identification in general, but also in attempting to combine a given multimedia file with similar multimedia content for the creation of computer generated answers.

These ranking and answering tools directly correlate to Q&A question forwarding research, which will need to begin ranking and identifying multimedia for its purposes of finding the correct potential answerers to questions. These tools are also necessary for searching archived questions, a vital component in reducing the number of duplicate questions. The problem is that they are focused on repurposing existing online content, not categorizing and ranking multimedia questions. Our system's implementation provides groundwork for the identification of multimedia content in a Q&A system through not only the association of question text with its multimedia, but also through the usefulness of specific multimedia content as determined by the question popularity. This covers both the ranking (question popularity) and answering (associated text content) challenges of MMQA systems.

In terms of Q&A security, Lee *et al.* [57] presented a study performed on Naver Mobile Q&A, a Korean Q&A website. Naver is similar to Yahoo! Answers, but with the added feature of allowing questions to contain images. The authors found that users often revealed personal details in their questions, such as place of birth and residence, health information, and place of work. The authors were then able to piece this information together for 2 of the users, even with only their small sample set of 20 users with 200 posted questions each. Their study was limited in that even though the website allowed for images, they only studied the text of the questions and answers. Additional risks that they did not examine were the contents of and data contained in these images, and risks to the Naver website itself from this multimedia. Additionally, another multimedia risk is that of copyright infringement and piracy. Chanmin *et al.* [72] explained that as multimedia use is growing, current copyright laws cannot keep up. They suggested a new network protocol that transmits information regarding the owner of the content that can aid law enforcement in prosecuting lawbreakers.

Given these privacy and security goals, the last step is finding which goals are actually being met. To do this, Herrmann [52] suggested that organizations create metrics to measure how well they protect the privacy of users. The metrics themselves would vary

widely based on the chosen application, but a metric allows an organization to ensure that changes or even day-to-day operations are not compromising privacy. Herrmann also referenced several principles from the Organization for Economic Cooperation and Development (OECD) [17], the most relevant being the Collection Limitation Principle. This attempts to set limits on what personal information should be collected, restricting it to what is required and no more. Additionally, she suggested, any information that is required should abide by the Purpose Specification Principle; before any data is collected, users should be told up front "what is being collected, how it will be used, to whom it will be disseminated, and how long it will be retained."

2.4 Summary

In this chapter, we have examined current solutions for Q&A systems, how they work, and what the latest technologies are. We have discussed much of the research that has been done in the field of Q&A, including question quality and its specified topics of question forwarding and multimedia classification. We have also previewed studies performed on Q&A security, along with their shortcomings. Within each of these topics, we have also detailed how this thesis contributes to that category of research. The following chapter will provide a detailed description of our Q&A system, how we used it to conduct our research experiment, and what specific security and privacy risks we chose to examine.

Chapter 3

The Experiment and Study

The purpose of this chapter is to provide a brief overview of our Q&A system and how it was utilized to conduct our research experiment, as well as a description of the security and privacy risks that were studied. We will first demonstrate our Q&A system, survey its features, and show what makes this website unique in the field of Q&A. We will then take a detailed look at the experiment itself and the processes that were used. Finally, we will review the security and privacy risks that we examined and preview our findings.

3.1 MultiQuery

The current version of MultiQuery was begun in the Fall of 2013, and has seen much improvement over the last three semesters of work. This website is programmed using the Ruby on Rails (RoR) framework with the model-view-controller (MVC) pattern [7]. This MVC architecture makes it is very simple to expand the system and add new functionality. Instead of requiring a large amount of code, a new website component or webpage can be added with only three or fewer simple files. By using modern technologies such as RoR and MVC, MultiQuery is better able keep pace with current technological trends and adapt to changing needs.

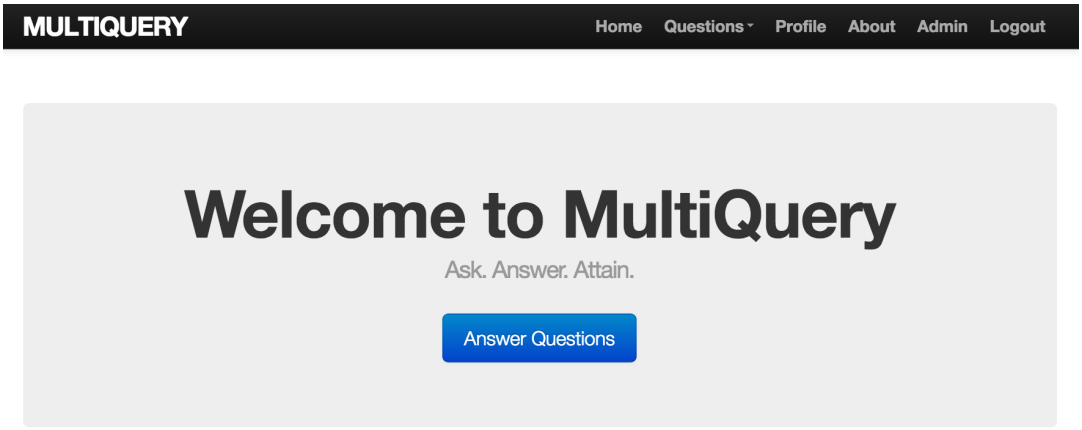

Figure 3.1: MultiQuery landing page design

Our Q&A system was designed using a combination of features from the Yahoo! Answers and StackOverflow websites. Similar to most other Q&A systems, MultiQuery allows users to create their own account, select a profile picture, etc. Users are also asked to select their favorite categories of questions, as well as categories that they consider themselves to be "experts" in. Once users register and log in, they are presented with a very simple landing page, as shown in Figure 3.1. Through the "Find User" link, users are able to find other users and view their personal profiles. In order to build community, users are encouraged to mark others as their "friend", or "follow" them. When that befriended or followed user asks a question, it will be forwarded to the friend or follower. This enhances the overall experience by increasing user interaction and getting questions to those who are most likely to invest time in answering them.

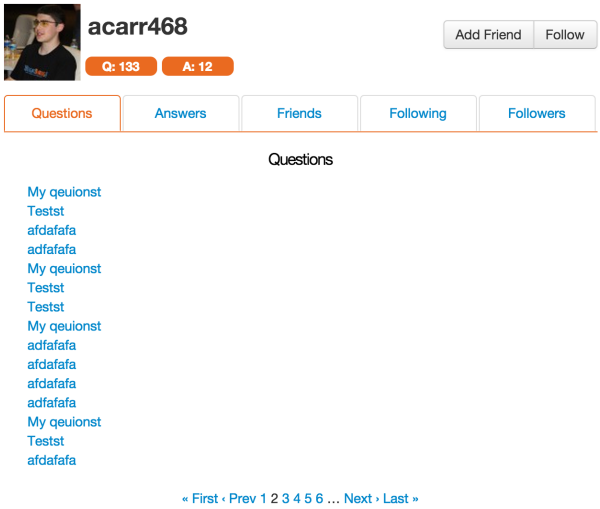

Figure 3.2: An example MultiQuery user profile

As seen through the example user profile in Figure 3.2, these profile pages show a great deal of information about the user. This information includes all public questions that they have asked and provided answers to, all friend relationships with other users, all users that they have followed, and all users that have followed them. Results are paginated into groups of 15 to minimize clutter. In terms of statistics, we provide the number of questions that a user has asked and answered in order to show their overall activity level.

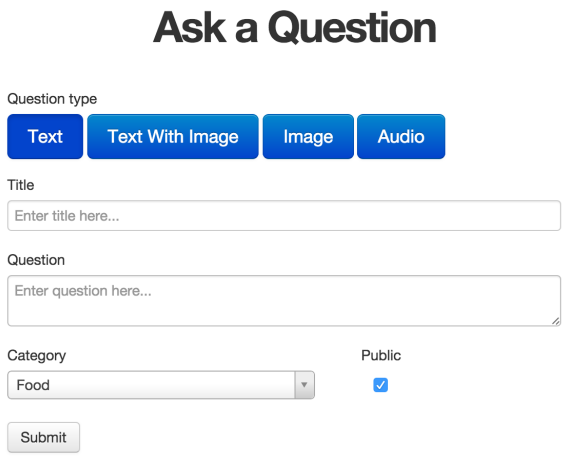

Figure 3.3: MultiQuery new question form

The other of the two landing page links (Figure 3.1), "Ask Question", directs the user to a webpage from which they can ask a new question. They select one of the four question types, either text-only, text with image, image-only, or audio-only, and then fill in all required information as shown in Figure 3.3. Other users are then able to view the question and all posted answers, and are also able to contribute their own answer (see Figure 3.4). All answers can be given an up or down rating by other users to indicate their opinion regarding that answer's quality. Finally, if this posted question was marked as "public", it is also included in a list of all available public questions, as seen in Figure 3.5.

3.2 The Experiment

The purpose of this experiment was to examine the relative effectiveness of multimedia in a question for receiving correct and/or helpful answers. There are many different combinations of multimedia that can be used in asking a question, and we chose the following: (1) A traditional question with plain-text only (the baseline), (2) A plain-text only

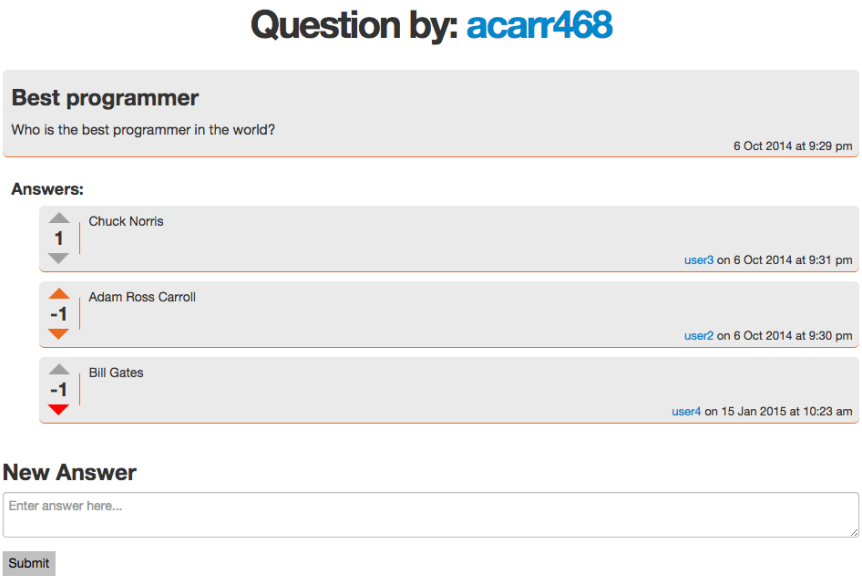

Figure 3.4: An example question on MultiQuery

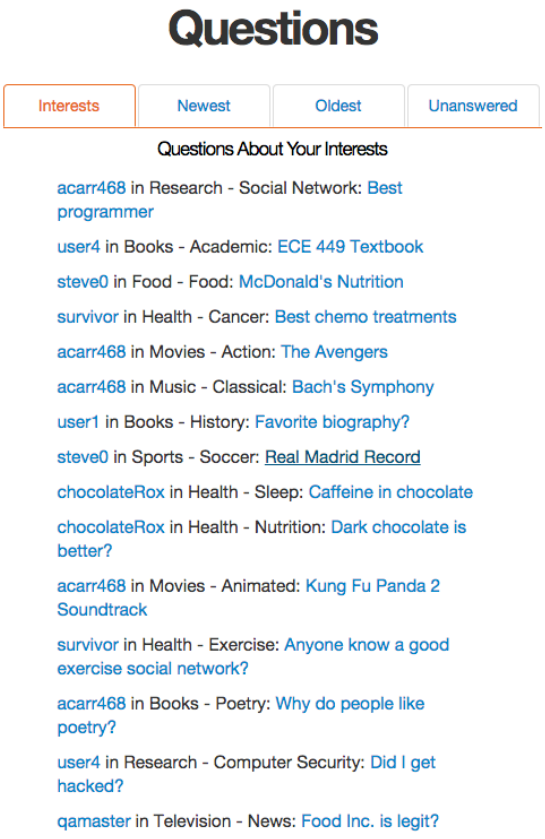

Figure 3.5: A list of questions on MultiQuery

question with the addition of an image, (3) A question containing only an image, (4) A question recorded as an audio clip by the asker. Our first aim is to find how the addition of images and audio to a question affects the correctness of answers, but then to also determine which of these forms of multimedia is the most effective. The goal of including type (3) is to discern if an image is beneficial only as support, as in type (2), or if the image itself is all that is required for an effective question. The goal of (4) is to determine if hearing a question aids in understanding through inflections and other vocal characteristics, or if it is hindered through added complications in comprehending the question. These questions were classified into 5 common categories: Food, Movies/TV, Music, Sports, and Technology. Creating these categories allowed us to answer our fourth research question, what categories are the most suited for multimedia Q&A.

In order to conduct the experiment, volunteers were sought among fellow classmates, friends, and family members. They were randomly placed into 1 of 4 groups, with an equal number of participants in each group, and assumed to have at least fundamental knowledge in each of the 5 categories. These groups were each assigned a number between 1 and 4. The multimedia questions would be assigned every week for them to answer, with all 4 multimedia types and all 5 categories used equally every week. Each of these questions would be systematically assigned to each group in such a way that throughout the course of the experiment, all participants would answer a relatively equal number of each type and category of question. This ensured that no group would bias the results of a question type or category through a heavier contribution to its data. A custom question assignment algorithm was written specifically for this, and is included in Appendix B.

Every week, 5 unique questions were written, and each of these questions were posted under each of the 4 different question types, while using the same textual question content between the types. This resulted in 20 total questions per week, which were evenly assigned across the 4 groups of users. Once answerers had submitted their answers, the answers would each be rated by study participants and marked as “correct” or “incorrect”. These rating assignments were performed by the participants instead of the researchers in

order to avoid any bias that we might impose. The assignments rotated for each group every week according to another custom algorithm. This rotation ensured a relatively even span of ratings from one group across all other groups, and no group was ever assigned to rate itself. This again prevented any group from biasing the ratings of another.

Participants were instructed to assess each answer and provide a score between 0 and 10, with 0 being a low rating and 10 high. Please note that other than the overall 0-10 scale, no other metric was specified to the raters. This was purposeful, as it gave raters the freedom to create their own metrics and use them consistently. In this experiment, it would be very difficult to write and communicate an objective metric to be used by all raters. To avoid this difficulty, we allowed each rater to fairly rate each answer relative to other answers. Because raters were not shown the user's name or any other identifiable information about the answerer, there could be no bias in their rating for or against a friend. The purpose of this experiment is not to quantify ratings, but to instead compare answer ratings between each type of multimedia question. Raters were also asked to indicate if an answer was correct or incorrect. Answer ratings from each type of the same question will be compared to determine which types of multimedia were the most effective in receiving good answers for that question. This analysis will be performed for each multimedia type overall (comparing the answer ratings and correctness percentages between each multimedia type), as well as for multimedia types under each question category (comparing the answer ratings and correctness percentages of each question category between each multimedia type).

To create these questions each week, archived text-only questions from our Q&A system were found and directly copied or slightly modified to meet the format of a multimedia Q&A website. Because of the large number of questions required, we also authored some of our own questions based on other questions that were found when browsing several different trivia websites [1, 2, 3, 4]. These websites contained many useful general-knowledge trivia questions about each of our question categories, which we copied or slightly modified for our own experiment. The goal was to keep questions short and concise to avoid excess work for study participants. Once the questions were chosen, we used Google to search for images

that could possibly aid in answering the questions, and we recorded audio mp3 files of the question texts being read aloud. Participants had never been shown any of these questions before, and the questions were chosen based on their ability to use multimedia. Anyone using a Q&A website specialized toward multimedia would seek to use said multimedia content, so these questions would be the type best fit for our system.

The week's questions were then posted and assigned to the corresponding groups who would be answering them by using our round robin question assignment algorithm. Each question was posted with its associated multimedia, along with a very short plain-text "subject" to categorize the question. Please see Figures 3.6 through 3.9 on the following page for an example of the four question types, with Figure 3.6 demonstrating a text-only question, Figure 3.7 text with image, Figure 3.8 image-only, and Figure 3.9 audio-only. Along with the question itself, additional data containing the correct answer(s) was attached to each question that would later be used by those rating the answers.

These assignments were performed for 10 weeks, with 50 unique questions and 200 total questions as a result of using these 50 questions for each of the 4 multimedia types. We also received 400 posted answers (2 per question), and 800 answer ratings (2 per answer). Please see Appendix A for a full listing of the unique questions that were used. All participants completed their assignments on time each week, and we acquired all desired data successfully. In order to study user preferences, participants also completed a survey prior to and following the experiment in which they were asked about their preferred and best-performing question multimedia types. This will aid us in answering our research question regarding what multimedia type an answerer preferred that the asker use. Finally, at the conclusion of the 10 week experiment, we posted an additional question using the text with image type. The catch here was that we used a **misleading** image to see if it would adversely affect the answers. This question asked about a prominent athlete's primary sport, but showed him in the uniform of another sport. A user who truly knew the answer would not be confused, but users basing their answer on the image instead of their own knowledge would provide an incorrect answer.

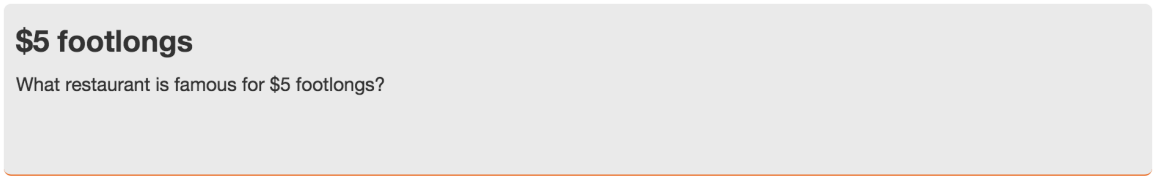

Figure 3.6: MultiQuery text-only question example

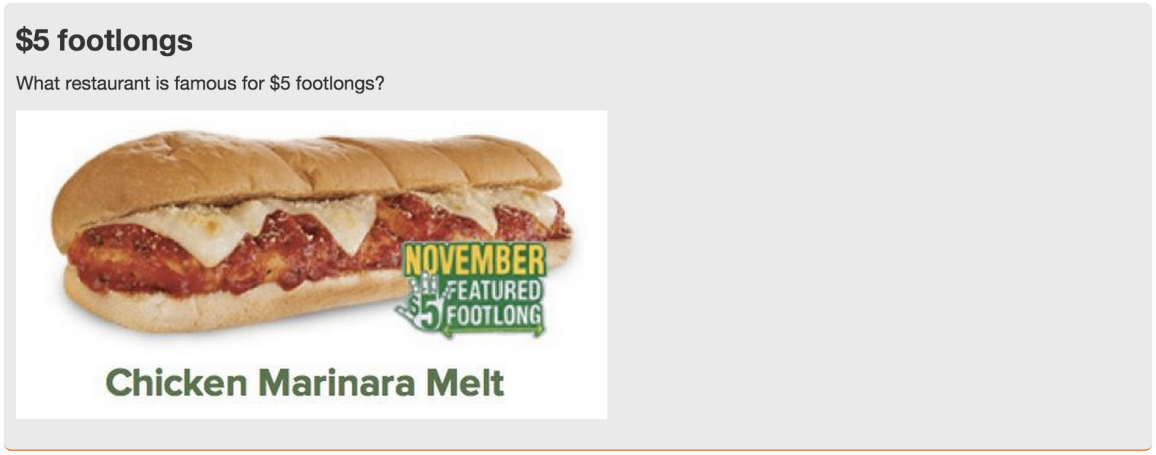

Figure 3.7: MultiQuery text with image question example

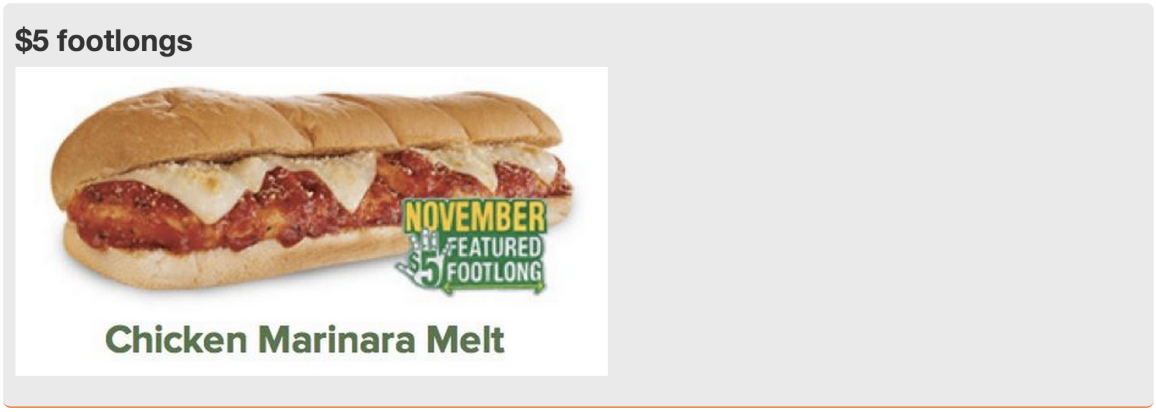

Figure 3.8: MultiQuery image-only question example

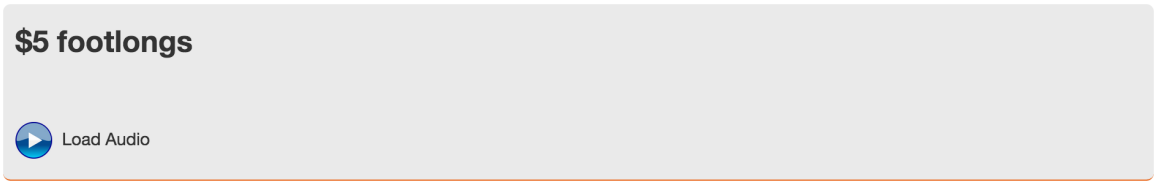

Figure 3.9: MultiQuery audio-only question example

Week Number	*38*		*39*		*40*		*41*		*42*		*43*	
Type - Category	***R***	***A***	***R***	***A***	***R***	***A***	***R***	***A***	***R***	***A***	***R***	***A***
Text - Food	x	**3**	**4**	4	2	**1**	**4**	2	3	3	1	4
Text - Movies/TV	x	4	1	1	3	2	1	**3**	**4**	4	2	1
Text - Music	x	1	2	**2**	**4**	3	2	4	1	1	3	2
Text - Sports	x	2	3	3	1	4	3	1	2	**2**	**4**	3
Text - Technology	x	3	4	4	2	1	4	2	3	3	1	4
Image - Food	x	4	1	1	3	2	1	3	4	4	2	1
Image - Movies/TV	x	1	2	2	4	3	2	4	1	1	3	2
Image - Music	x	2	3	3	1	4	3	1	2	2	4	3
Image - Sports	x	3	4	4	2	1	4	2	3	3	1	4
Image - Technology	x	4	1	1	3	2	1	3	4	4	2	1
Text/Image - Food	x	1	2	2	4	3	2	4	1	1	3	2
Text/Image - Movies/TV	x	2	3	3	1	4	3	1	2	2	4	3
Text/Image - Music	x	3	4	4	2	1	4	2	3	3	1	4
Text/Image - Sports	x	4	1	1	3	2	1	3	4	4	2	1
Text/Image - Technology	x	1	2	2	4	3	2	4	1	1	3	2
Audio - Food	x	2	3	3	1	4	3	1	2	2	4	3
Audio - Movies/TV	x	3	4	4	2	1	4	2	3	3	1	4
Audio - Music	x	4	1	1	3	2	1	3	4	4	2	1
Audio - Sports	x	**1**	2	**2**	4	**3**	2	**4**	1	**1**	3	**2**
Audio - Technology	x	2	3	3	1	4	3	1	2	2	4	3

Table 3.1: Algorithm results from the first six weeks. Columns "R" and "A" refer to the group numbers that are "Rating" and "Answering" the question respectively.

In order to assign participants questions to answer and answers to rate, two custom algorithms were designed. When assigning participants questions to answer, the most important feature was that the assignments span all 4 multimedia types and all 5 question categories as evenly as possible during the 10 week experiment. This would ensure that no group of participants would heavily influence the results of any combination of multimedia type and question category. For the rating assignments, it was vital that each group rate all other groups the same number of times, while also preventing them from being asked to rate their own answers. This again was to ensure an even influence on the data, and to prevent groups from giving themselves a biased score. We implemented these algorithms into our system in the form of round robin assignment based on the current week number in the year. Please see Appendix B for a more thorough explanation of our algorithms.

Table 3.1 shows a sample of results from using these algorithms over the first 6 weeks of the experiment (our experiment began on the 38th week of the year). For each week, the user group number that was assigned to each question is listed under the Rater and Answerer columns. Some values for group #4 near the top of the table are highlighted to demonstrate that raters never rate themselves. Group #4 was first assigned to rate group #3, then #2 the following week, then #1, and the cycle repeats. Similarly, values for the "Audio - Sports" question are highlighted to show that a different group is assigned to answer that specific question type and category each week, starting with group #1 the first week, group #2 the second, group #3 the third, and group #4 the fourth, before cycling to group #1 again. This table empirically demonstrates the correctness of our algorithms.

Our Q&A system provides an excellent platform to conduct Q&A research, and underwent thorough testing before approval. Despite its robustness, additional work was necessary in order to administer our experiment, such as user restrictions and administrative tools. Since users were assigned specific questions each week and not allowed to view those of others, access to assigned questions needed to be tightly controlled. To do this, a database table called QuestionViewer was created for each question that was assigned to a user. A user could only view a question for which they had a QuestionViewer relationship, and they were not able to post multiple answers or see the answers of others. Similarly, a QuestionRater database table was created for each question to which a user was assigned as a rater. Raters were not shown any information about the answerer in order to ensure that there was no bias.

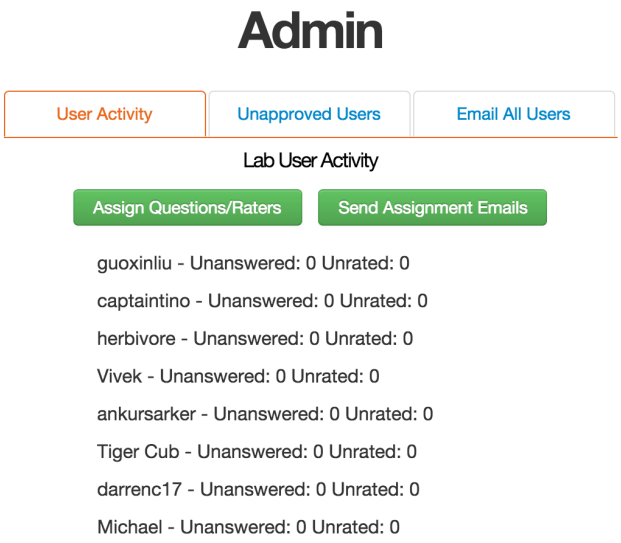

Figure 3.10: The MultiQuery admin panel

To administer the experiment, we also built an admin tool that automatically runs the answering and rating assignment algorithms via a web interface (see Figure 3.10). Once questions have been posted, a single button click executes the algorithms, and a second button click sends an email informing the participants about their assignments for that week. From this page, we were also able to quickly determine if a participant had any answer or rating assignments that had not yet been completed.

At the conclusion of the experiment, we examined the average answer ratings and correctness percentages for each multimedia question type overall, as well as for each specific category of question. We also compared the average answer ratings and correctness percentages for each user overall and for each specific category, attempting to determine to what degree multimedia performance was dependent on the users themselves. Finally, we examined the user survey results, their trends, and how these results compared to our observed data. The results of these evaluations will be detailed in the following chapter.

3.3 Multimedia Q&A Security

In addition to the experiment, we also conducted a thorough examination into image and audio risks, as well as user and webmaster security dangers for multimedia Q&A systems. We first investigated risks from image metadata and facial recognition, then voice recognition, and finally risks to the webmaster from the uploaded multimedia content.

When any photo is captured by a device, information (called metadata) is stored alongside the actual pixels of the image in a format called Exif [20]. There are industry standards for Exif, with NISO MIX being the most common [5]. The actual values contained in this data vary based on the type of camera that was used, and may be absent altogether, but common elements include the number of pixels, date that the photo was taken, and the camera model (see Figure 3.11). All of this information is potentially incriminating, and can be used by an attacker to exploit known vulnerabilities in a specific camera device. In conducting our research, we evaluated the Exif information for several images taken with

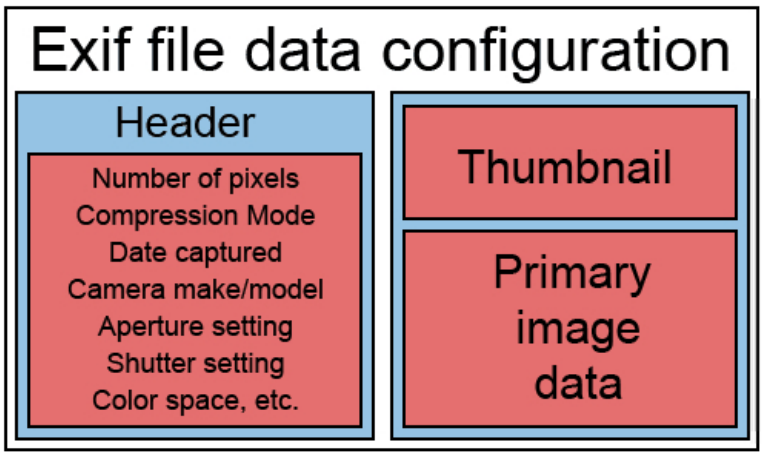

Figure 3.11: Standard format for Exif [20]

the camera of an Apple iPhone 5s, using an app called ExifViewer [22]. We found many potential vulnerabilities, which will be detailed in the following chapter. We also studied potential risks for facial recognition, examining two governmental programs that focused on this new technology.

In addition to images, MultiQuery, Ask.fm, and some other Q&A systems support audio or video questions and answers, and these usually contain recorded human voices. There are privacy risks associated with uploading these audio files, specifically though voice recognition. We studied the possibilities for voice recognition and what risks might arise through this technology. Additionally, on the security front, we examined current copyright and privacy protection laws for users. Once content is posted online, it can be very difficult to control its use.

Finally, we examined the risks to us as multimedia Q&A webmasters. Not only are there risks to a Q&A website's users through this multimedia content, there are also risks for the Q&A website owners themselves. When users are given the ability to upload any image or audio file that they wish, this opens the door for questionable activity. Users may upload material that is illegal (copyrighted music, images, etc.), malicious (viruses, online scams, etc.), or explicit (pornography, sexual content, etc.). Webmasters need to be aware of these risks and implement adequate safeguards.

3.4 Summary

In this chapter, we have have shown the design of our Q&A system, our experiment and its goals, algorithms developed for the experiment, and potential security and privacy risks to multimedia Q&A websites. The experiment itself was a success in that it acquired all desired data with high levels of user participation. The security study was likewise successful, as we found many security holes, but also many possible solutions. In the following chapter, we will examine the results of the experiment, surveys, and security study, with our conclusions and recommendations to follow.

Chapter 4

Results and Analysis

The purpose of this chapter is to analyze the data that was gathered through the experiment and security study, explain the implications of the findings, and provide suggestions for current and future Q&A systems. For the experiment, we begin with a high-level analysis of the overall rating and correctness percentages across the questions themselves, and then take a more detailed look at the users and how they were affected by the multimedia. Next, we provide the results from our security study, before finally revisiting and providing answers to our research questions that were discussed in Chapter 1. In preview, we found that (1) multimedia is a very useful tool for receiving correct answers, especially for the text with image and audio-only question types, (2) Sports questions received great benefit from the addition of multimedia, but Movies/TV did not, (3) users strongly preferred the text with image type and performed at their best when using it, (4) while higher performing users were aided by questions containing multimedia, lower performing users were actually worse when multimedia was added, (5) category experts often received lower ratings and correctness percentages when using multimedia, but higher values when answering questions outside of their expertise, (6) multimedia confused many users on the "trick" question when a misleading image was used, (7) many risks exist to Q&A website users, especially through image metadata and voice recognition, and (8) there are many risks to Q&A website owners from illegal, malicious, and explicit content.

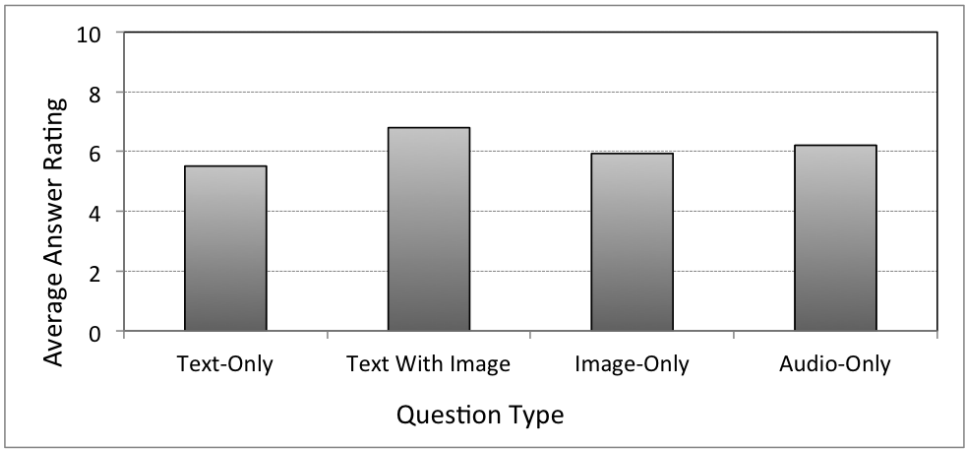

Figure 4.1: Average answer rating for each type of multimedia question

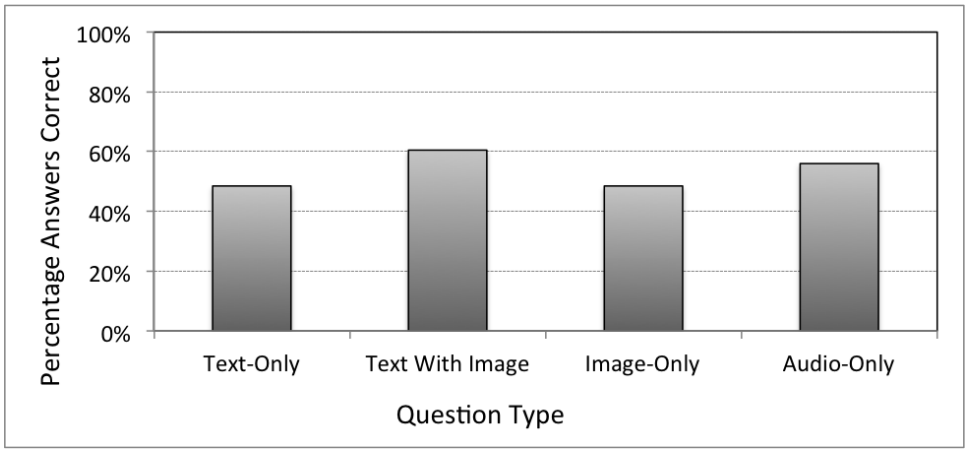

Figure 4.2: Correct answer percentage for each type of multimedia question

4.1 Multimedia Question Analysis

To begin the analysis, we examined the average answer ratings and percentage of answers marked as correct for each type of multimedia question. We found that the highest rated answers were received under the text with image type with an average rating of 6.8, and the lowest from the traditional, text-only type with an average rating of 5.5 (see Figures 4.1 and 4.2). As mentioned earlier, these answer ratings were valued within a 10 point scale, and all following figures of this type will be using the same scale on the Y axis. These scores reveal a 24% higher rating for text with image over its more traditional

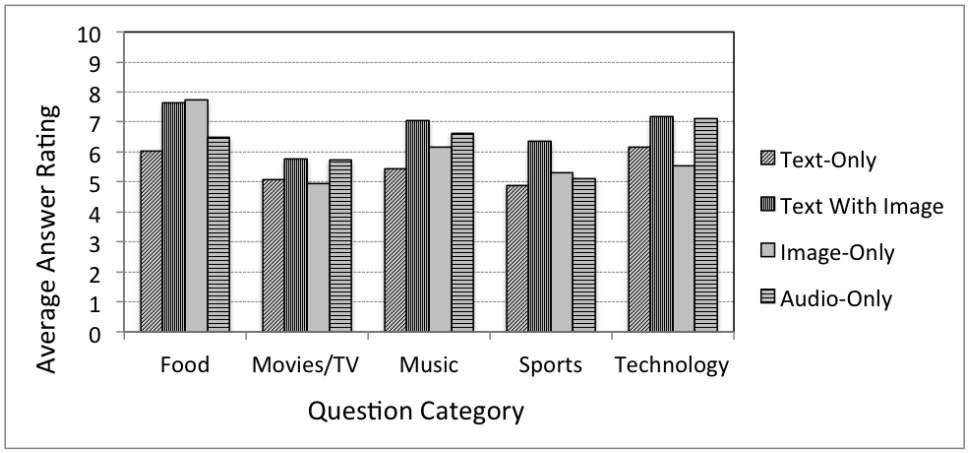

Figure 4.3: Average answer rating for each multimedia question type per category

counterpart. Similarly, the highest correct answer percentage was also received under the text with image type at 60.5%, and the lowest tied between text-only and image-only at 48.5%; a 25% increase for the text with image type. Surprisingly, audio-only performed 13% and 15% better than the text-only baseline in the average rating and correctness scores respectively, even though it was the same text being recorded. This indicates that the addition of multimedia is very helpful overall, but we must dive deeper and see how the addition of multimedia affected each individual category of question.

As Figure 4.3 shows, the Food and Sports categories received the highest answer ratings when using images in general, while Movies/TV, Music, and Technology performed best with text with image and audio-only. Movies/TV was a tough area, receiving very little benefit from the addition of any of the multimedia types. This, in contrast to the Sports category, which showed a 30% increase in average rating between the highest rated multimedia type, text with image, and the text-only baseline. One possible reason for this is that questions in Movies/TV, Music, and Technology often require more information within the question, and the image-only and text-only types are not as well suited to providing this. Food and Sports, on the other hand, tend to be more straightforward and rarely require much more information than what an image can provide.

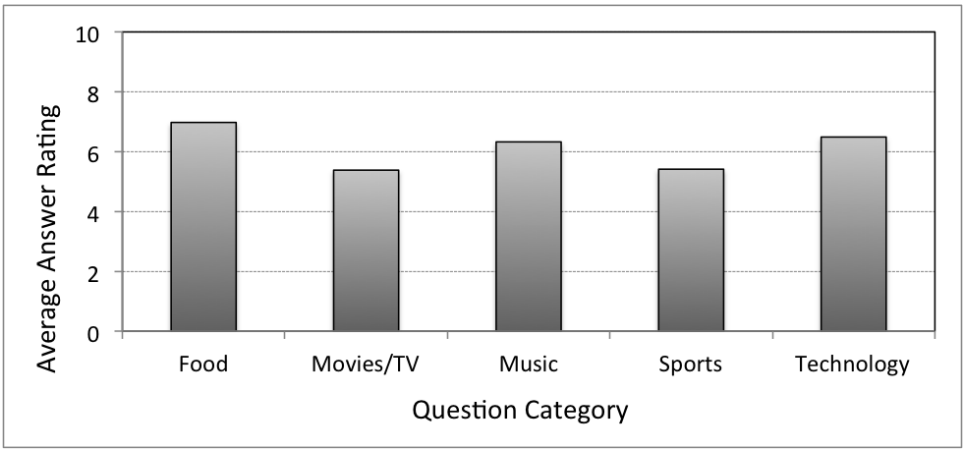

Figure 4.4: Overall average answer rating per question category

In looking at each specific multimedia type, while image-only was very helpful in Food, it was even worse than the traditional text-only question type for both the Movies/TV and Technology categories. A probable reason for this is that Food questions are more straightforward (as mentioned above), identifying a vegetable for example, while Movies/TV have many questions about movie information or the actors themselves, and a simple image is not able to convey the necessary details. Audio-only was more consistent, always performing above the text-only baseline. Audio-only contained the same text as the text-only, so it is understandable that any benefits or drawbacks from this form of multimedia would be consistent across all categories. Text with image was the strongest contender, always or almost the highest rated in every category, as expected after reviewing Figures 4.1 and 4.2. We suggest that this is due to the greater amount of information available, including not only the full question text, but also a supplementary image, as opposed to only text, recorded text, or the simple image.

Figure 4.4 provides an overall average for ratings across each category, with some categories faring better than others. Questions in the Food category, for example, were among the highest rated overall, and the challenging Movies/TV questions among the lowest. Through this and the previous figure, we can see that the text with image multimedia type was always at or near the top in average answer rating for all categories, and the text-

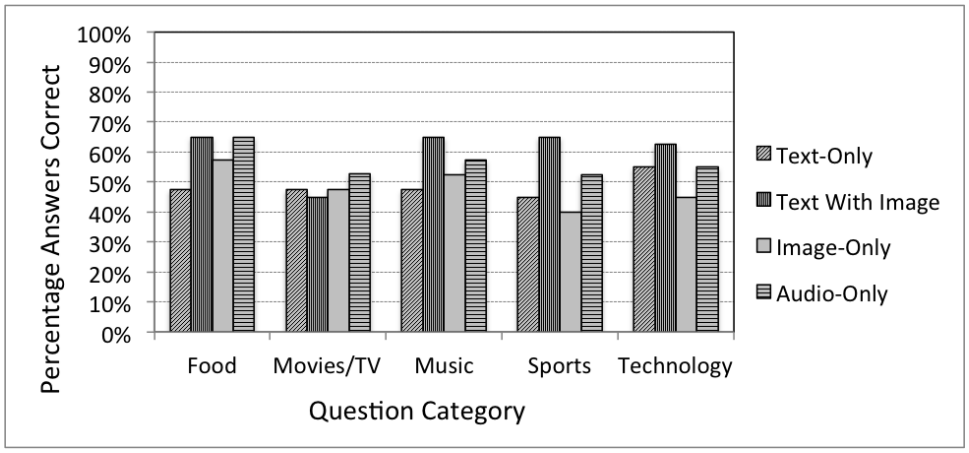

Figure 4.5: Correct answer percentage for each multimedia question type per category

only type consistently, and almost always, among the lowest. We can also observe that Food questions appear very well suited to a multimedia Q&A system, while Movies/TV do not receive much benefit. Possible reasons for these results are that, as mentioned above, many Food questions are attempting to identify an object such as a vegetable, while Movies/TV require more detailed information that is less easily transmitted via an image.

We can see similar patterns when looking at the percentage of answers marked as correct in Figure 4.5, with the Y axis representing the average percentage of answers marked as correct for each category (0-100%). All similar figures of this type will be using the same 0-100% scale on the Y axis. The text with image type took a strong lead in all but the Movies/TV category, where it was actually the worst of all question types. This closely resembles our answer rating findings in Figures 4.3 and 4.4, and we suggest for the same reasons. Correctness is arguably more important than the overall rating, as a correct answer is the ultimate goal, so this figure should be given significant attention. These results indicate that overall, a text-based question with the addition of an image almost always receives higher rated and more often correct answers than a traditional text-only question. Audio-only questions, although not as highly rated as text with image, also showed higher ratings and correctness percentages than text-only in all categories. Disappointingly, image-only questions saw mixed results; helpful in some categories such as Food, but unhelpful

in others such as Sports and Technology. This indicates that images by themselves are not consistent tools for asking questions, but should only be used as a supplement.

In this section, we have seen through the text with image multimedia type that adding an image to a question can greatly aid the average asker in receiving the correct answers. Images alone are not enough, and in some cases less helpful than the traditional text-only questions. The biggest surprise to us was how well the audio-only questions performed relative to text-only. Despite the fact that it was the exact same text in the audio recording, the audio-only questions consistently received better answers. We suggest that this is because users pay more attention to the audio-only questions due to their novelty, and as a result devote more energy into answering them. Now that we have examined results from the perspective of the questions, we will take a look at the users themselves and how each responded to this new concept of multimedia questions.

4.2 User Survey Results

The purpose of this section is to examine the results of surveys sent to participants prior to and following the experiment. In both surveys, users were asked which of the 4 multimedia types they felt would/did receive their best answers, and which types they would/did prefer to answer. For a Q&A system to be effective, users must be willing and ideally excited to use its features. When introducing multimedia to this type of system, it must be done in a way that users will want to use the new multimedia features. The survey's goal was to not only document user preferences, but also to show if those preferences changed as they used the website and learned how to utilize multimedia in a question. Table 4.1 contains the survey results, and the following paragraphs will explain the implications.

In the first survey, all but one of the users predicted that the text with image type would receive their best answers, and all but two predicted that the text with image type would be their favorite to answer (outliers bolded for emphasis). User #8 was one of the latter two users, and they responded with text-only as their answering preference because it

Period	*Prior To Survey*		*After Survey*	
Metric	*Best Answers*	*Preferred*	*Best Answer*	*Preferred*
User 1	Text w/Image	Text w/Image	Text w/Image	Text w/Image
User 2	**Text-only**	**Text-only**	Text w/Image	Text w/Image
User 3	Text w/Image	Text w/Image	Text w/Image	Text w/Image
User 4	Text w/Image	Text w/Image	Text w/Image	Text w/Image
User 5	Text w/Image	Text w/Image	**Text-only**	Text w/Image
User 6	Text w/Image	Text w/Image	Text w/Image	Text w/Image
User 7	Text w/Image	Text w/Image	Text w/Image	Text w/Image
User 8	Text w/Image	**Text-only**	Text w/Image	Text w/Image

Table 4.1: Results from the user surveys taken prior to and following the experiment, predicting/reviewing which question type will/did receive their best answers, and which question type they will/did prefer to answer. The outliers are bolded for emphasis.

"is the fastest kind of question for me to answer." This participant raises an important point: even if multimedia aids users in answering a question, every bit of additional information potentially increases the overall time required to answer that question. Audio-only questions especially, as answerers must listen to the entire question instead of simply skimming its text. While the average adult English reading speed is 300 words per minute, the average speaking speed is half of that; 150 words per minute [24, 27]. This requires the user to spend twice as long receiving the question. Overall, though, users were very much in favor of adding images to a traditional text-only question prior to the experiment.

After the experiment, **all** users were convinced that the text with image type was the best, and 100% of the users also listed it as their preferred question type to answer. This indicates that the initial concerns from User #8 were unfounded, and the benefits gained versus the additional time required are well worth it. And the unanimous user prediction that their best answers had been given under the text with image type was mostly accurate, as seen earlier in this chapter. The fact that 100% of the users were aware of their (perceived) increased answering ability for this question type further proves the type's effectiveness. In answering our research questions, it also shows that this preference, being unanimous, is not strongly dependent on the specific user.

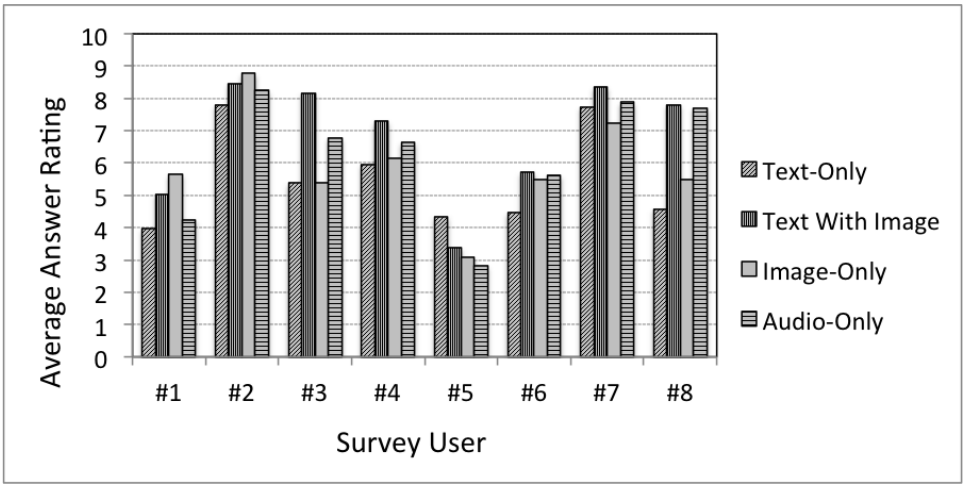

Figure 4.6: Average answer rating for each multimedia question type per user

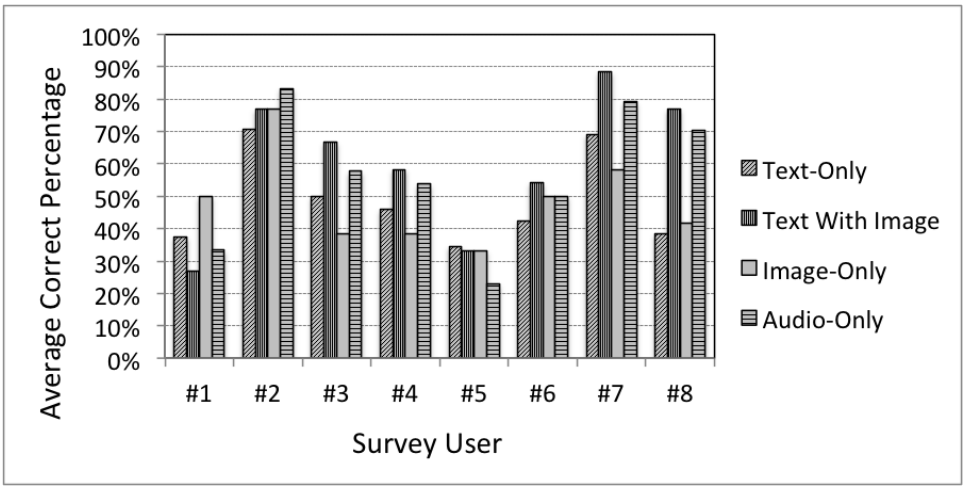

Figure 4.7: Correct answer percentage for each multimedia question type per user

4.3 User Analysis

Now that we have seen the preferences of users, we will examine how each user actually performed across each multimedia type and category. This analysis will show how consistent users are across each type, and to what extent multimedia enhances a user's ability to answer questions in their selected area of expertise.

First, we will examine the average answer ratings and percentage of answers marked as correct for each multimedia type among the users. As Figures 4.6 and 4.7 show, some

users performed significantly better than others. User #2, the highest performing user, received both an answer rating and correct answer percentage almost 250% higher than that of User #5, the lowest performing user. Most of the users' results show trends similar to Figures 4.1 and 4.2, with text-only receiving the lowest rating, and text with image the highest. This demonstrates that the results were not skewed by a few select users, but instead reflect a consistent trend. The anomalies were among the lower performing users, especially User #5. It appears that users with a lower ability to answer questions are more confused than aided by multimedia. This may be because of difficulty in understanding the asker's accent in the audio questions, as some users were from countries outside of the USA and questions were recorded by an American asker. Or, another possibility is that these foreign users experienced difficulty in identifying unknown objects from other cultures in the image questions. In contrast, the top users showed a significant level of improvement when answering questions containing multimedia, obviously able to understand the additional content. User #7, for example, answered a staggering 89% of their text with image questions correctly, versus only 69% of their text-only questions. Overall, it is evident that top performing expert users are aided by multimedia, while lower performing are not. Because these lower performing users did not answer questions well for any of the question types, even text-only, their answers will be less often correct or accepted by a Q&A system's users. Improving the more often accepted answers of the top performing users is much more important, and that was in fact accomplished by the addition of multimedia.

Among those that excelled, however, each user showed a different level of improvement across each multimedia type. Some users performing better using text with image, and some audio-only. This seems to indicate that different users are better suited to answering questions of differing multimedia types. These two multimedia types should therefore be considered as independent parts of expert identification systems, and also then made a part of question forwarding algorithms.

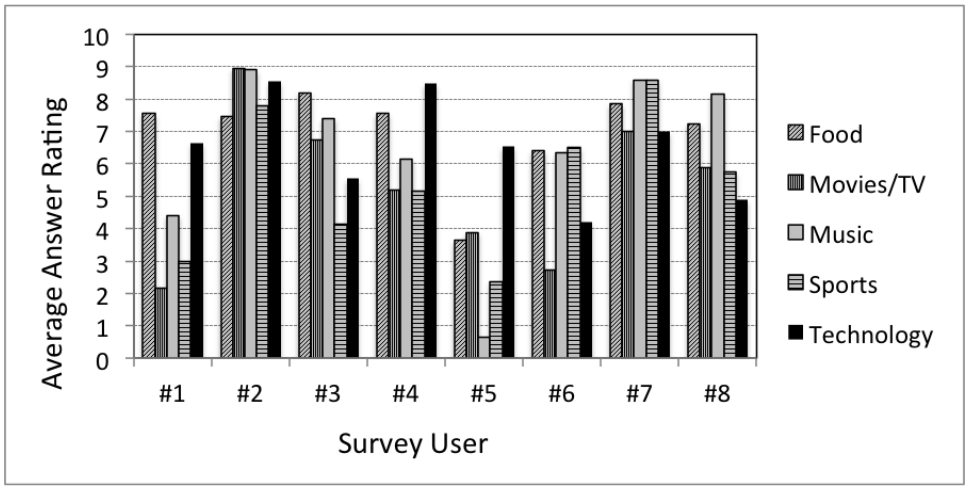

Figure 4.8: Average answer rating for each multimedia question category per user

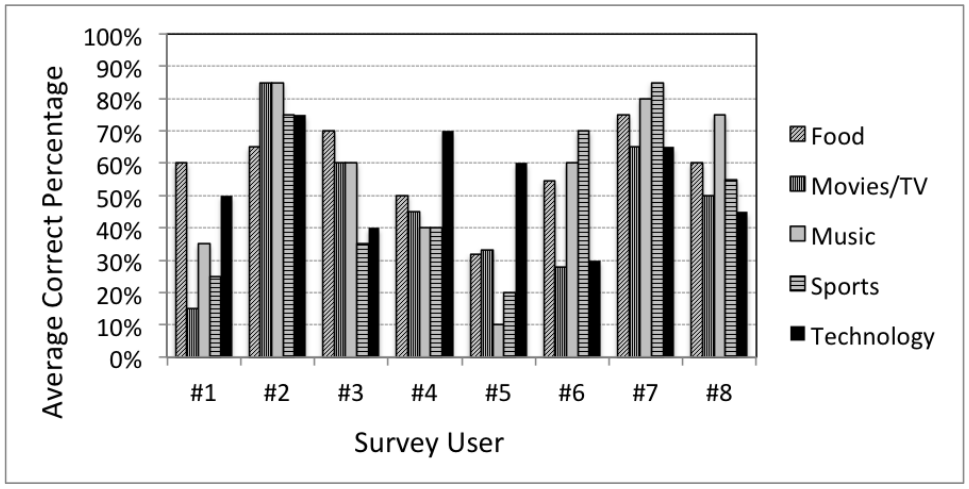

Figure 4.9: Correct answer percentage for each multimedia question category per user

Next, we will examine how users performed individually across each of the 5 question categories. As seen in Figure 4.8, showing the average answer rating for each multimedia type per user, and 4.9, showing the average correct answer rating for each multimedia type per user, each user showed a wide variety of question answering abilities. Of the 8 users, 2 users scored highest in Food, 1 in Movies/TV, 1 in Music, 2 in Sports, and 2 in Technology. This indicates that all categories of questions are useful to users, and could be considered as part of a multimedia Q&A website. Additionally, it should be noted that those rated as the top performing users received fairly consistent ratings across all categories. While

this may be expected, it negates the notion that the top users simply excelled in a few of the categories and were mediocre in answering others. On the contrary, these users seem to demonstrate that a multimedia website having only a few knowledgeable or expert users can effectively provide answers for a wide range of question categories.

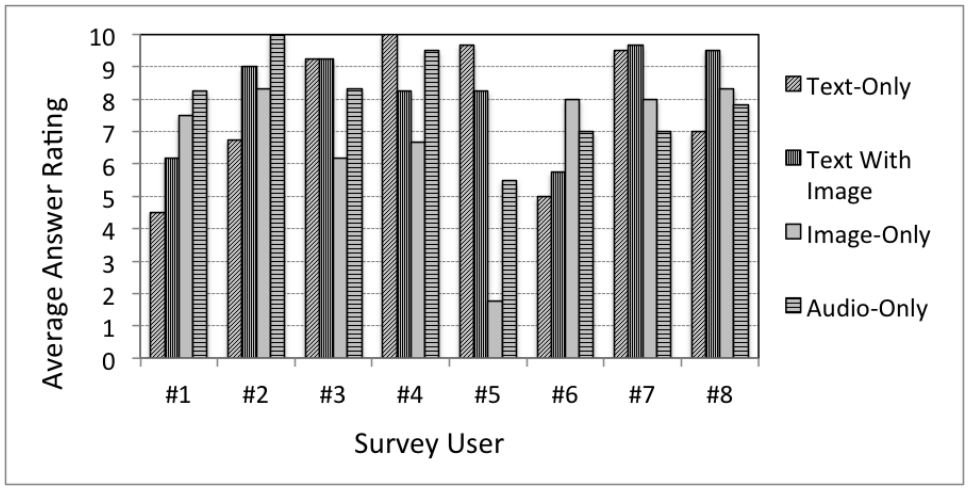

Figure 4.10: Average answer rating in expert category for each multimedia question type per user

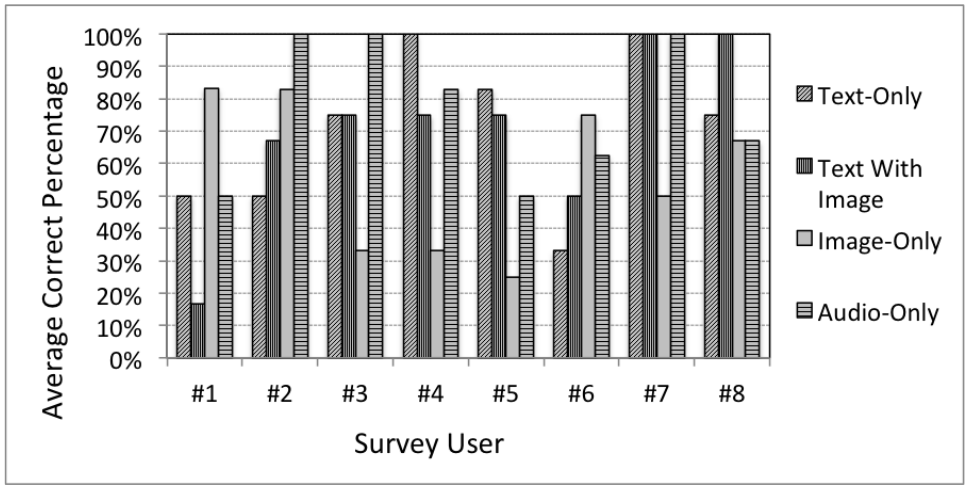

Figure 4.11: Correct answer percentage in expert category for each multimedia question type per user

Prior to the study, users were asked to select the category that they considered themselves to be the greatest experts in. Of the 8 users, only 5 correctly predicted their best category, although the other 3 users were very close. But within that expert category, how did each user perform when using multimedia? Figure 4.10 shows how each user's average answer rating compared across each multimedia type for their selected expert category, and the results are surprising. 3 of the 8 users actually scored the highest using the text-only question type, with no multimedia used at all. Similarly, Figure 4.11 shows that 3 of the 8 users had their highest correct answer percentages with the text-only type. This shows us that multimedia does not have as much of a helpful effect on users who are already experts in a topic, but has a much greater benefit when answering questions outside of their expertise. We suggest that this is because expert users require very little (if any) additional information in order to answer a general-knowledge question inside their area of expertise. This seems to indicate that multimedia is best suited to a generalized Q&A system containing many categories of questions such as Yahoo! Answers, and not something more specialized such as StackOverflow with a very limited range of detailed categories.

4.4 The "Trick" Question

In addition to the other experiment assignments, we also asked a final, "trick" question to all groups at the conclusion of the experiment. Using the text with image type and the Sports category, we posed the question in Figure 4.12, "What primary sport did John Elway Play?", but used a misleading image. The purpose of this was to determine if users would answer the question by simply basing it on the image itself, not their own knowledge. The correct answer was "football", but the image was of the athlete in a baseball uniform, and 75% of the users submitted an answer of "baseball". *This indicates that image questions are a double-edged sword. They can be very helpful if applied correctly, but almost guarantee incorrect answers when chosen poorly.* This may be due to the trusting nature of users on a Q&A system, assuming all content to be accurate and truthful. But this

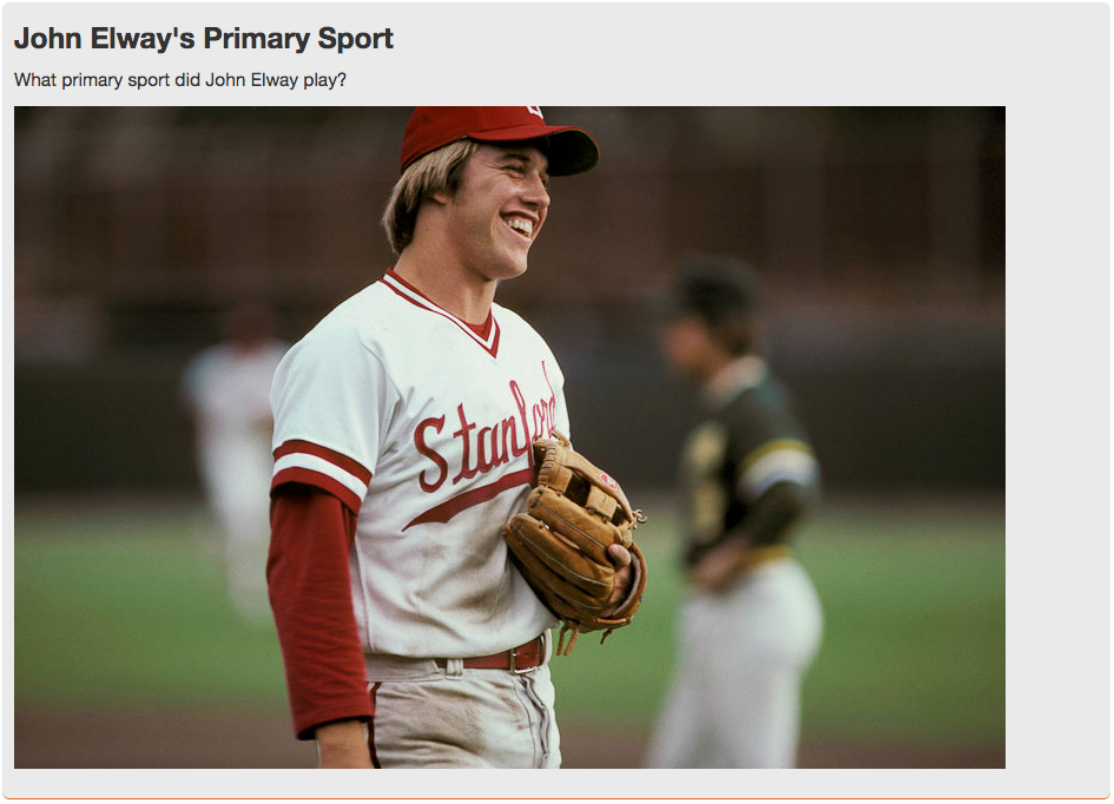

Figure 4.12: "Trick" question using the text with image type

confusion and possibility for poorly chosen content is almost impossible to avoid, as it is up to the asker to supply the image content. Askers should be advised to be careful in what image(s) they select, as their choice of image could have an adverse affect when naively chosen.

4.5 User Privacy and Security Risks

In addition to the experiment, we conducted a thorough examination into image, audio, and webmaster security and privacy risks for our multimedia Q&A system. Our findings, implications, and solutions from this study are in the following sections, followed by the applications in Chapter 5.

4.5.1 Image Risks

As mentioned in Chapter 3, we captured several images with an Apple iPhone 5s in order to study security implications from the attached metadata. From these images, we were able to determine the version of the Apple iOS operating system that the phone was running (via the "camera software" field), as well as the resolution of the image (see Figure 4.13). An attacker could potentially use this data to determine not only what OS vulnerabilities were available to exploit, but also possibly predict the model of phone based on the resolution of the camera and other camera Exif fields.

In addition to the hardware information, almost all mobile phones and even some traditional point-and-shoot cameras will attach GPS coordinates (location data) of where the image was taken [13]. In our tests, one of the same photos used above was traced to a Walmart Supercenter (see Figures 4.14 and 4.15). This information, combined with the time that the photo was captured, could be used to trace an asker's shopping, working, or even travel habits over time. Or, even more concerning for parents, reveal where their child attends school. As a result of this tracking, it would be very simple for a malicious user to determine the location of the asker's home, place of work, etc. From that, it could also be determined when an asker is away from home, especially when on a vacation or other extended trip. This would allow a thief to know when a home is uninhabited and an easy target for robbery or other mischief.

Another common element in photos posted online is that they often contain faces of either the uploader or their friends. These faces may seem anonymous and untraceable, while in reality they are anything but. As mentioned in the introduction, there are many algorithms that can determine the identity of a face with remarkable accuracy. In a simple case, face matching, Facebook's DeepFace algorithms are able to determine if two faces are of the same person within a 97.25% accuracy [71]. Facebook uses this algorithm when a user uploads a new photo in an attempt to determine if any of the faces in the photo are of that user or their Facebook friends. But there is nothing to stop this algorithm (or others) from searching an entire database of users instead of just a small group of acquaintances.

Image width
3264 pixels

Image height
2448 pixels

Date/time
2014:10:10 21:11:23

Date/time original
2014:10:10 21:11:23

Date/time digitized
2014:10:10 21:11:23

Camera software
8.0.2

Artist
Not set

Figure 4.13: Phone and camera information from ExifViewer

GPS latitude
34.908420

GPS latitude reference
North

GPS longitude
82.332275

GPS longitude reference
West

GPS altitude
297.153061

GPS altitude reference
Above sea level

Exposure time
1/30 seconds

Figure 4.14: GPS coordinate information from ExifViewer

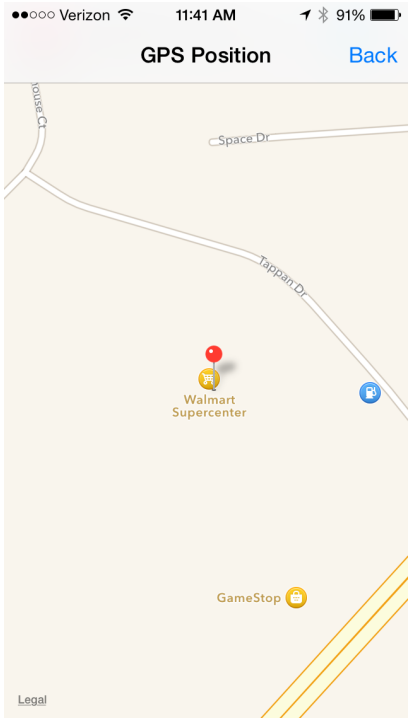

Figure 4.15: GPS coordinates map trace using ExifViewer

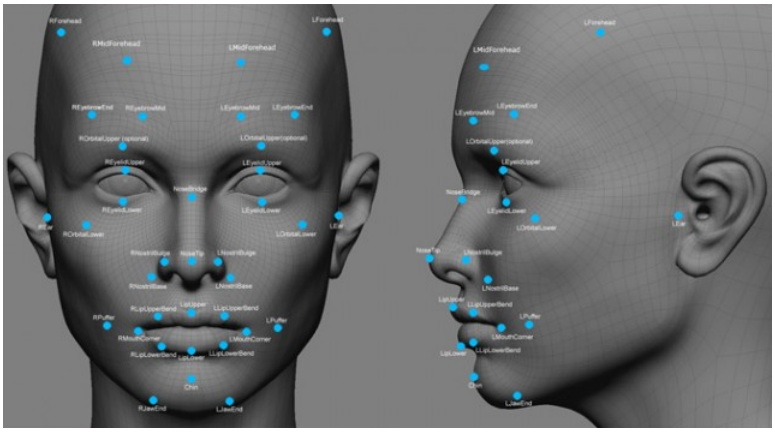

Figure 4.16: DeepFace identifies key facial features [71]

Possibly even more disconcerting, governments are developing their own databases and algorithms for facial recognition. In the US, the Federal Bureau of Investigation (FBI) has launched the Next Generation Identification program that will further enhance their current database of fingerprints with the additional storage of recognized faces [16, 32]. These faces will be used for investigations and other crime related incidents. This plan is nearing completion, and should be close to operative, if not already [28]. Any images uploaded to a Q&A website have the potential to one day be taken and used by the FBI in these databases.

Another US group, the National Security Agency (NSA), was recently discovered to have been collecting millions of photos per day from intercepted private communications across the globe. Of these, a staggering 55,000 per day were of a high enough quality for facial recognition [34]. While the NSA claims that these photos are mainly captured from the communications of non-US citizens and used only for counterterrorism and other security purposes, the algorithms and growing databases of faces have the potential of being repurposed later if laws change. Almost anything posted to the Internet is accessible by the NSA's wiretaps given the proper permissions, making any uploaded image a potential privacy risk, including those uploaded to a Q&A website.

4.5.2 Audio Risks

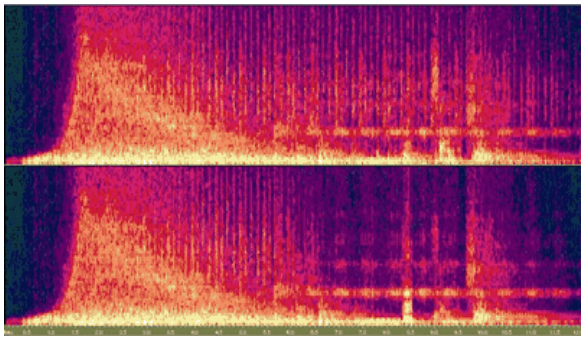

Figure 4.17: A comparison of two voiceprints [42]

In addition to image risks, we found that audio questions have many of their own potential risks, particularly through voice recognition. When uploading a recorded audio clip of their voice, an asker would probably expect their friends or acquaintances to be able to recognize them. This should be an immediate concern to any user, because if it is recognizable to anyone at all, it can be made recognizable to a computer. Every human has different acoustic features in their speech. When their voice is recorded and digitally displayed, this is known as a "voiceprint" [31]. Much like a fingerprint, a voiceprint is unique to the speaker [68]. If an asker's voiceprint is known, a match can be determined via "speaker verification" [40]. All that this requires is two audio recordings, one from a database or other known source, and the other from the uploaded file contained in the question (see Figure 4.17 for a voiceprint example). And as detailed in the following section, no file uploaded to the Internet is private. Therefore, any user who uploads an audio clip cannot be guaranteed anonymity from the government or anyone else. This has direct implications on our audio-only question type, as these will always consist of a user recording their own voice and uploading it to our system. Any user who asks this type of question will need to be aware of the risks, and possibly advised to obscure their voice in some way if they have any privacy concerns.

4.5.3 Loss of Ownership

Finally, another group of risks we found of uploading multimedia content online is that the user can no longer fully control its access and use. MultiQuery has a "friend" filter that allows askers to only share questions with users that they have labeled as their friends, but even this is a risk. As Frank Nagle and Lisa Singh described, users are often more

than willing to "friend" a possibly malicious user and divulge information if they have a "mutual friend" connecting them [69]. If this "friend" is truly malicious, any content that they upload is potentially public.

Another risk is piracy. While the uploader usually maintains all copyrights [50], Internet piracy is easy to do and difficult to track. The uploaded photos can be used by anyone else online for whatever purposes they desire with little chance of facing legal action. To make this even trickier, many social networks and photo sharing websites including Facebook and Picasa actually claim a royalty-free license on all uploaded content, including images [51]. This ownership allows them to share or otherwise use all uploaded images any way that they deem fit.

Even uploads to websites that allow for "private" questions can be made public through security breaches. A recent breach of note was that of Apple's iCloud cloud storage service in August of 2014 [55]. Several celebrities' accounts were broken into, and copies of their private photos, many explicit in nature, were shared and made public. Any Internet upload or storage service, no matter how secure, can be attacked. While no legitimate Q&A system should be hosting photos of this nature, it demonstrates that any multimedia content uploaded for a question, even those with restricted private or "friend-only" access, can one day become public.

4.5.4 Webmaster Security Risks

4.5.4.1 Illegal Content

In addition to user risks, we also studied risks to ourselves as Q&A webmasters. One major gray area of the modern Internet is piracy, the sharing of files that are copyrighted or otherwise not intended to be made freely available. A webmaster, by allowing for the upload of images and audio files, is giving an outlet for a user to upload restricted files that should not be shared. This includes not only copyrighted files, but also confidential and otherwise restricted content. Webmasters who host these files are responsible not only for removing them when requested, but in some cases may also be held liable for the copyright

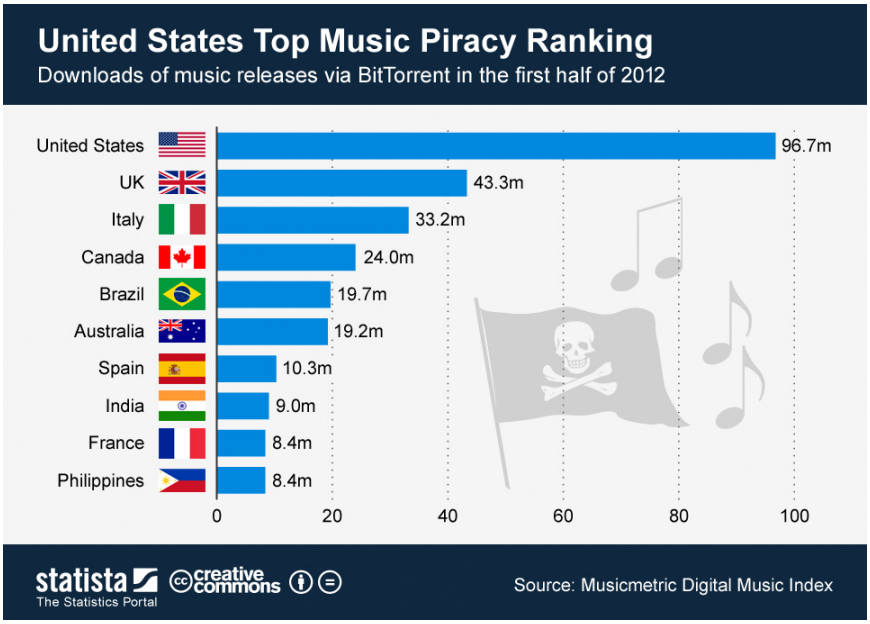

Figure 4.18: Global music piracy downloads for 2012 [76]

infringement of the uploaders [25]. We found that the Digital Millennium Copyright Act (DMCA) is designed to protect webmasters from this third-party copyright infringement if they meet certain requirements, creating a "safe harbor" for the webmasters [10]. These requirements include things such as a posted copyright policy, quick removal of the offending files when reported, and an agent to deal with these reports [11]. This sounds simple until the website grows, requests pile up, and the "takedown agent" becomes a full time job. If these requirements are not met, the webmaster is no longer protected by the law. And piracy cases, even small ones, can lead to significant penalties. For a recent case in 2012, one woman was found guilty and fined $222,000 for the sharing of only 24 songs [29]. And compared to the overall number of pirated music files worldwide, this case was not even significant (see Figure 4.18 for worldwide levels of piracy). These fines are unaffordable for anyone, particularly webmasters and small companies on a tight budget.

4.5.4.2 Malicious Content

Another danger for Q&A webmasters is malicious content. This content may be malicious toward users, the website itself, or both. In terms of danger to users, it is possible to upload files that plant viruses or run other forms of malicious code on any machine that downloads that file. For example, on websites that display images with a simple HTML "img" tag, it is possible to execute PHP scripts or other malicious content instead of an actual image [73]. Audio files, in certain formats such as Adobe Flash, can also contain exploits that link users to a malicious website when the files are opened [9].

Content can also be malicious toward the website itself. In terms of storage, simply the large size of an uploaded file could be its own form of attack. While we may expect a "normal" user to ask questions with images that are a few megabytes in size, the uploads could be much larger. Askers could also upload files that are not truly images, but simply have an image file extension. These large files tax the server's bandwidth and can take up valuable system space, particularly for smaller companies. Finally, we found that the web form for uploading may be a security threat itself if implemented incorrectly. Bogdan Calin of Acunetix describes several ways that the form can be exploited, one example being a common, simple PHP script that accepts the uploaded file via a POST request and copies it to a temporary location on the sever [43]. This copied file, if malicious, could potentially compromise the entire server onto which it was placed.

4.5.4.3 Explicit Content

A final danger to Q&A webmasters that we found is explicit or sexual content. When allowing users to upload any image or audio file that they wish, this opens the door for pornographic or otherwise objectionable images, as well as vulgar audio files. While some Q&A websites may embrace this type of content, the vast majority, including our system, are seeking to avoid the stigma of being an "adult" website and wish to be open to all ages. Pornography, estimated to account for 30% of global Internet traffic [30], can be a lucrative business. But some forms of pornography are from people looking to spread their content

simply for someone's embarrassment. This, otherwise known as "revenge porn" [33], should be a direct concern for any Q&A website that allows for image questions. Allowing this or vulgar "adult" audio would narrow the market of the website and must be prevented.

4.5.5 Solutions to these Risks

Once we discovered these risks, our next goal was to find methods to prevent them. The "problem" of multimedia will never go away, but only get worse. As seen earlier in this thesis, one of the greatest sources of multimedia is from sharing via smartphones. Cisco published a white paper in 2014 predicting the average monthly global data traffic from smartphones over the next 4 years, and it is growing exponentially (see Figure 4.19). As this continues to grow, our problems and dangers will grow along with it. There are two primary areas to investigate that can protect users and webmasters from these dangers: website implementation and website moderation.

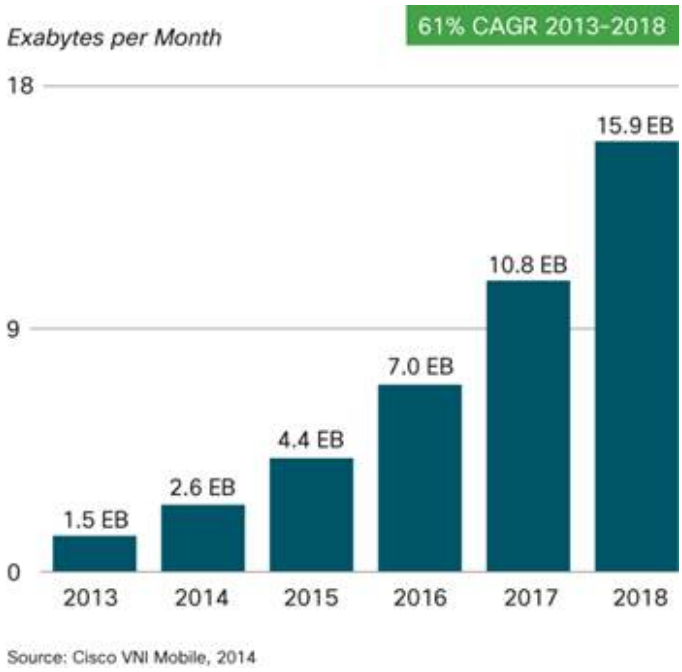

Figure 4.19: Worldwide mobile data use per year [45]

4.5.5.1 Website Implementation

When any website is built, developers can mitigate or even prevent security risks simply through their methods of implementation. One risk mentioned in Section 4.5.1 was image metadata, which could possibly allow a malicious user to track the asker or exploit vulnerabilities of their camera device. One way of avoiding this is to simply remove

all metadata from an uploaded image. This is not always a good option because it can be useful to have the metadata in the future, some websites such as Flickr keep this information for users that want to prove image authenticity, etc. [48]. But for Q&A websites such as ours, the benefits are not worth the risks. A downside to stripping metadata is that some processing is required on the image, but this is only a small price to pay for the privacy gains. This is a specific choice that a webmaster needs to make given the website's goals, and then needs to inform users of that choice.

Another risk that was mentioned is how the website actually displays data, particularly images (Section 4.5.4.2). If an asker uploads a PHP script or other disguised file instead of an image, a webmaster who naively displays the image via a simple "img" HTML tag while not checking the file type will put users at risk. Instead, the webmaster must ensure that the uploaded file is truly an image, and add filters in their content management system to reject any files that do not match the specific image criteria they are expecting [73]. In similar fashion, file sizes should also be checked and rejected if they exceed a specific threshold, unique to the website and its market. None of these security measures currently exist on MultiQuery, but may be addressed through future work.

4.5.5.2 Website Moderation

Some risks cannot be prevented beforehand and must be remedied after the fact through moderation. The crucial problem faced here is for explicit content. While filters already exist that screen textual content, there is no way to detect if a multimedia file is explicit or otherwise objectionable before it is uploaded. Because of this, Q&A moderators must be available to look for and remove the offending files. Copyrighted material, on the other hand, is not as easy to moderate. Moderators cannot be expected to know if any given content is a breach of copyright or not, so they must rely on knowledgeable users to report these files before any action can be taken. And as mentioned in Section 4.5.4, this copyright moderation is mandated by law. These moderation tasks may even require a large team of moderators, depending on the size of the website.

4.6 Answering the Research Questions

Now that we have thoroughly examined the experiment and security study, we will summarize that information here and apply it to our research questions from Chapter 1. Overall, it is evident that text questions with the addition of images were the most effective in receiving both good and correct answers in the experiment. Recorded audio questions came in at a close second, with the text-only and image-only question types rounding out the bottom. Image-only questions were volatile, sometimes performing excellently and other times horribly. We also discovered many privacy and security risks to our users and ourselves, including image metadata, facial and voice recognition, and illegal, malicious, and explicit content, as well as solutions to mitigate or even prevent these risks. Below, we will present the answer for each of our research questions.

1. **How does the use of images or audio (multimedia) in a question affect the correctness of answers?**

   As mentioned above and seen earlier in the chapter, the addition of images to a text question resulted in a 25% increase in correct answers, and the use of audio a 15% increase. This shows images to be an incredible help in receiving good answers, and audio showed promising performance as well. Questions containing only images, on the other hand, received the same number of correct answers as the text-only questions. This demonstrates that images by themselves are not as useful as might be thought, but are much more effective when used as a supplementary tool. *Conclusion: this experiment shows that both images and audio, when used effectively, are helpful in the search for answers and would be well worth the additional effort.*

2. **Between images and audio, which of these two multimedia types is more effective in receiving correct answers?**

   As hinted at in the previous question, images, when applied as supplements instead of replacements, are the most effective overall in receiving correct answers. This

type had the highest ratings and correctness percentages overall, and the highest for almost every category. *Conclusion: it is clear that the addition of images results in the greatest chance of receiving the correct answer, with even better results than audio. This type would be the best first choice for a Q&A system.*

3. **When answering a question, what forms of multimedia (if any) do answerers prefer the asker to be using, and to what degree is this dependent on the specific answerer?**

   As the survey results showed in Table 4.1, all but two users preferred the text with image question type before the experiment, and **all** preferred it by the end of the experiment. One of the initial holdouts noted that answering questions with additional multimedia content required a greater level of effort, and therefore would not be as preferable to answer. This raised a very important point: the benefits of multimedia as explained in research question #1 must be counterbalanced by user willingness to accept the extra work, no matter how small, in analyzing these additional multimedia question components. Whether it be an additional image component, or a time consuming audio component, all of these multimedia types require differing levels of effort from the answerer. *Conclusion: participants unanimously preferred the somewhat complicated yet more effective text with image question type, showing that the additional effort is well worth the reward. This question type also received the highest answer ratings in the experiment, further demonstrating this preference.*

4. **What categories of question (if any) are the most suited to a multimedia Q&A website?**

   Figure 4.5 indicated that some categories fared better than others, with Sports showing the most impressive improvement through the addition of images. Other categories, such as Movies/TV showed minimal if not negative results from the addition of multimedia. Overall, it was found that questions in the Food category were the most effective questions asked in general, but Sports and Music clearly saw the great-

est improvement from the use of multimedia and would therefore be the most suited to a multimedia Q&A website. A Q&A website focusing on these categories seems to be a good decision. *Conclusion: some categories can take great advantage of multimedia, such as Sports, but others such as Movies/TV, will have greater difficulty.*

5. **What are the security and privacy risks involved with a multimedia Q&A website?**

   As we have seen, privacy and security risks abound, especially for images. Images contain metadata, and this often holds private information about the user's location and camera device. In addition, images and audio have facial and voice recognition risks respectively, which may allow for the identification of the asker or friends. Finally, webmasters face risks from illegal, malicious, and explicit uploaded content, and must provide appropriate safeguards. We found several solutions, particularly through the removal of image metadata, safer upload forms, and a stronger presence of moderators. *Conclusion: Images and audio all contain potentially identifiable information that should be removed or somehow disguised. Additionally, webmasters must safeguard and police their Q&A website to avoid security risks from the uploaded content.*

4.7 Summary

In summary, we have found that (1) multimedia is a very useful tool in receiving correct answers, with the addition of images to a text-only question receiving a 25% increase in correct answers, and use of audio a 15% increase, (2) Sports questions received great benefit from the addition of multimedia, but Movies/TV did not, (3) users strongly preferred the text with image type, unanimously selecting it as their favorite by the end of the experiment, (4) users that performed well in general performed even better with the addition of multimedia, but those that performed poorly overall were worse with the addition of multimedia, (5) few users performed better in their topic of expertise when multimedia was added, (6) when faced with a challenging, confusing multimedia question of the text

with image type, 75% of users answered based on the incorrect image and not their true knowledge, (7) many risks exist to Q&A website users through location and camera device information contained in uploaded images, facial and voice recognition in uploaded images and audio, and potential loss of private information through piracy or security breaches, and (8) there are many risks to Q&A website owners through illegal content and associated copyright laws, malicious content that targets the Q&A system or users, and explicit content that deters widespread user adoption of the Q&A system. Now that we have observed all of our experimental data and security study results, we will summarize and conclude this thesis in the following chapter.

Chapter 5

Conclusions and Discussion

This chapter summarizes and concludes our analysis of the Q&A multimedia experiment and security study, and also proposes additional features that we would like to add to future versions of our Q&A system. Finally, we suggest Q&A research topics that should be studied further, now knowing some of the fundamental principles behind the different uses of each multimedia type.

5.1 Conclusion

In review, this thesis has provided the following contributions:

1. We have designed and implemented a multimedia Q&A system, a website designed specifically for our experiment. It allows users to ask and answer questions, rate questions, follow other users, etc. Built using the Ruby on Rails framework, this is a very maintainable and expandable website. And, while many Q&A systems allow for image or audio questions exclusively, we allow for the posting of text, image, and audio questions all on the same system.

2. We have collected and analyzed the data received from the experiment, including the average answer rating and correctness percentage for each multimedia type and question category, both overall and for each specific user, the users' preferred

and perceived best multimedia type to answer, what topics of questions are best suited to each multimedia type, how well users answered multimedia questions in their chosen category of expertise, and how users respond to confusing questions.

3. We have provided both a discussion and analysis of multimedia security and privacy risks. Adding multimedia functionality to a Q&A website creates new risks that must be understood and addressed. These include privacy risks to users from images and audio, as well as security risks to website owners from illegal, malicious, and explicit user uploads.

Q&A systems have enabled users to share information through the asking and answering of questions online for many years, and multimedia is beginning to work its way into these systems. In order to study multimedia's effects, we first created a website named MultiQuery that allows users to ask and answer questions. Once this was created, we administered a 10 week study in which volunteers were asked 5 questions a week. These questions spanned multiple different categories, and utilized 4 different multimedia types: (1) A traditional question with plain-text only, (2) A plain-text only question with the addition of an image, (3) A question containing only an image, and (4) A question recorded as an audio clip. The goal of our study was to examine how each of these 4 types of multimedia affected the quality and correctness of the answers that were received.

Our system required several unique features in order to perform the experiment, including custom algorithms to assign the questions each week and a custom administrative panel for use by the researchers. The experiment performed very well, in that it received all of the desired information and was completed within the proposed time window. We found that the text with image multimedia question type received the most correct answers by far, with a 25% increase seen over the text-only type. Audio-only also saw an increase, with a 15% greater correct answer percentage over that of text-only. Image-only, however, saw mixed and inconsistent results. This indicates that the addition of multimedia is very helpful overall, but best used as a supplement or somehow still communicating literal text.

A counterpoint to the use of multimedia, as one participant noticed, was that it requires different amounts of time to evaluate each question element. Some questions, such as those with only an image, may be faster to view than a text-only question. But other types, such as audio-only, may take twice as long to review. While our participants were required to answer their assigned questions, a commercial Q&A website whose users are free to answer whatever they please may have difficulty with some multimedia types. Potential answerers may avoid the more complicated types, unwilling to invest the additional effort. Not only this, but any website that hosts multimedia content must account for the space required to store it and bandwidth used to display it. Our experiment's relatively short audio questions were roughly 100 KB in size, but longer questions would be much larger. Because of these added costs, modern Q&A websites may have to adjust their pricing structure, or charge a premium for some multimedia services.

Additionally, we made several observations regarding question categories. When multimedia was introduced, the Food category especially showed great improvement in question quality and correct answer percentage. Technology and Music also showed improvement, particularly for the text with image types. On the other hand, questions in the Movies/TV category saw mixed and often negative results from multimedia. While this was disappointing, the mixed nature of the results does not seem to indicate that multimedia should be avoided for this category. On the contrary, all categories received some sort of benefit from at least one of the multimedia types, the challenge is simply in determining which type is best for each category.

We also observed some confusion from the use of multimedia for a few users. While many users excelled and showed a great improvement when answering questions containing multimedia, those that performed poorly in all categories actually performed even worse with multimedia. This indicates that users with a lower ability to answer questions will be more confused than aided by the addition of multimedia. This confusion was also seen in the "trick" question that was asked, where a deceptive image was used alongside a legitimate question text. 75% of the users answered based simply on the image and not their own

knowledge, demonstrating this confusion. Among those that excelled, however, each user showed differing improvements across each multimedia type, with some users performing well using text with image, and some audio-only. This indicates that these two multimedia types should be integrated as independent parts of expert identification systems, and also then made a part of question forwarding algorithms. Additionally, users also performed, on average, worse on questions in their area of expertise that contained multimedia than those of the text-only type. This indicates that multimedia is a greater aid to non-experts, and is not as necessary for those that have a wider knowledge of the topic, which means that multimedia would be best suited to a more generalized, all-encompassing Q&A system.

Finally, we discussed the many security and privacy implications of introducing multimedia to a Q&A website, including metadata and facial recognition for images, voice recognition for audio, piracy and security breaches for users, and finally the security risks for webmasters. For MultiQuery and other Q&A systems, these risks can be addressed several ways. The primary methods are through more secure website implementation practices, as well as efficient and thorough moderation. Even so, some problems such as facial recognition cannot be fully solved. It is up to the asker to be diligent and selective in what they upload and share, and up to webmasters to keep that uploaded content private and secure from hackers and other malicious users. Given all of this information, we make the following recommendations:

1. **Q&A websites should support image questions, but require additional text**

   Our results revealed that the text with image type received a great increase in answer ratings and the percentage of correct answers received, but image-only did not. Multimedia's disadvantage of the additional time required to review the question appears minimal for images, making images the multimedia type of choice. These findings are supported by the users as well, who clearly supported the text with image type in their survey responses and higher levels of performance. Image sizes should be restricted to reduce bandwidth and storage costs, though, as mentioned in the previous chapter.

2. **Q&A websites should carefully consider audio-only questions before implementing them**

   While the audio-only questions saw an increase in answer ratings and the percentage of correct answers received, this increase came at a cost. We cited that English audio takes twice as long to review as does written text, and other languages would be similar. This means that users may skip over these questions, either due to impatience or simply not understanding the accent of the asker. Additionally, potential answerers may not currently be in a setting that allows for noise. Q&A websites should carefully consider these drawbacks, as well as the additional storage requirements, before embracing this type of multimedia.

3. **Q&A websites need to preserve user privacy by removing all image metadata, enhance the security of upload forms, and give greater effort toward moderation**

   While some websites may need this metadata, GPS coordinates and other, sensitive details are completely unnecessary for a simple Q&A website. Q&A websites should remove all of this associated data from an image before it is saved to the system's servers. Additionally, Q&A website owners must safeguard their systems through more secure upload forms and strong moderation.

5.2 Future Work

Our Q&A experiment performed excellently as a preliminary study into the effects of multimedia, but there is much work still to be done. We recommend several enhancements to future studies, including the addition of video-only and text with audio multimedia question types. This form of audio as a supplement to a text question was not included as part of the experiment, as we simply investigated audio-only, but it may be helpful in categories such as Music, when an asker is attempting to identify a song or musical piece. Videos were not a part of this study either, and would be a useful point of examination after

seeing the other Q&A apps such as Ask.fm. We also recommend a more formal statistical evaluation using hypotheses and establishing statistical significance. Finally, future studies would benefit from a larger user base over a longer time period. Our short term experiment provided an excellent exploratory study into the field of multimedia Q&A, but there is much still to learn through long term system analysis.

Much of the related work that we covered in Chapter 2 referenced question quality and forwarding algorithms, and these need to be enhanced based on our results. Question quality is heavily affected by the presence or absence of multimedia, some more than others. Current question quality equations must account for this, and provide higher or lower scores based on the presence of multimedia. Additionally, question forwarding algorithms must account for the fact that some users are more adept at answering different types of multimedia questions. A user's average score for each multimedia type should be combined with the existing equations when performing the calculations.

Finally, MultiQuery is in need of several security and privacy improvements. We need to not only implement a file type checker for our image uploads, but we also must also protect these images by removing their metadata. Additionally, we must strengthen our moderation practices to ensure that no illegal, malicious, or explicit content persists on our system.

In summary, we have seen great improvements in the helpfulness and correctness of answers received for questions that used multimedia, particularly in the text with image questions. We have also made several observations regarding the security and privacy concerns of these systems, and steps to safeguard against them. It will be interesting to take these preliminary results and apply them to each individual topic of Q&A research, such as question quality, question forwarding, and multimedia Q&A security.

Appendices

Appendix A Appendix: List of Questions From the Study

The following is the list of questions that were used in the experiment, grouped by question category. Most questions where of the "what" variety, attempting to identify an object or idea. Most also had fairly straightforward answers, but a few allowed for wider and a more comprehensive range of answers.

Food:

- What vegetable is long, green, and shaped kinda like a tube?
- What herb creates a cooling sensation when touched or eaten?
- What is maize?
- What is the sweet substance made by bees?
- What is the popular food used to carve jack-o-lanterns during Halloween?
- What is the topping made of ground beef and beans that often goes on hot dogs?
- What restaurant is famous for $5 footlongs?
- What food is chorizo?
- What is yogurt?
- Bundt is a type of what food?

Movies/TV:

- How are movies usually physically stored?
- What movie star was well known for his karate and sometimes wore a cowboy hat in his movies? He also has many memes about him on the internet.
- What is the movie The Guardians of the Galaxy based on?
- What actor played Han Solo in the Star Wars film series?
- What is the theme of the Bourne series of films?

- What series of TV comedies stars Lucille Ball who is married to a singer/actor named Ricky?
- What company founded by Steve Jobs and now owned by Disney is famous for creating animated movies, including Toy Story?
- What is the name of the new Stars Wars movie?
- What longstanding TV show asks people questions worth different dollar amounts?
- What genre of movies was John Wayne famous for?

Music:

- What is the instrument that is made out of brass and makes a really loud, high-pitched noise when hit together?
- What does an orchestra's conductor wave to keep time?
- What are common instruments played with Jazz music?
- What genre of music did Louis Armstrong play?
- What instrument is a big version of a guitar?
- What instrument has keys, pedals, and strings?
- What genre of music does Taylor Swift perform?
- When a marching band performs, what do other performers often carry and wave around while the band is performing?
- In music, what is syncopation?
- What is it called in music when you divide a beat into smaller, equal subdivisions?

Sports:

- In basketball, what is the big semicircle line around each hoop?
- What sport, similar to soccer, is played in water?
- What does the acronym NFL stand for in sports?

- What sports event is El Clasico?
- What is the mercy rule in baseball and other sports?
- What primary sport does Peyton Manning play?
- What sport has the Lou Gehrig Award?
- How long is a marathon?
- What is the new name of the college football Bowl Championship Series (BCS)?
- In golf, what is the name for the person that carries a golfer's clubs around?

Technology:

- How do you set up Apple pay?
- What is the numlock key on a computer keyboard for?
- What is Sony's primary (and large) game console called?
- What does the SQL stand for, referring to programming languages?
- When talking about computers, what does CPU stand for?
- What software sold by the company called Adobe is used to edit photos?
- What format of movies is on a disc like a DVD, but is newer and higher quality (high definition)?
- Why are Cortana and Siri usually female voices?
- In terms of the internet, what does WWW stand for?
- What is typically used on a laptop instead of a mouse?

Appendix B The Algorithms

Algorithm 1 Question Answering Assignments

```
1: procedure AssignQuestionAnswerers
2:     weekNumber ← current week number of the year
3:     questions[] ← this week's questions, sorted by multimedia type and then category
4:     for i := 0 to (questions.size() - 1) do
5:         groupNumber ← ((weekNumber + i) % 4) + 1
6:         group ← group with id of groupNumber
7:         Create a QuestionViewer for each member of group
8:     end for
9: end procedure
```

The pseudocode for assigning questions to answer is detailed in Algorithm 1. The aforementioned goal of assigning questions evenly is accomplished by rotating the starting group while still assigning questions in the same order. Each week, as *weekNumber* increases by 1, the starting *groupNumber* will also increase by 1. Because the *questions* array is sorted by multimedia type and category, and the types/categories of questions are the same every week, the assignment order for the question types/categories will also stay the same every week. With a different starting group number each week, then, each group will experience a 4 week rotating pattern of assignments. This results in groups answering each question type/category combination either 2 or 3 times, ensuring an even spread of the question types and categories among the participants.

Algorithm 2 Answer Rating Assignments

```
1: procedure AssignAnswerRaters
2:     weekNumber ← current week number of the year
3:     ratingRotation ← weekNumber % 3
4:     questions[] ← last week's questions, sorted by multimedia type and then category
5:     for i := 0 to (questions.size() - 1) do
6:         groupNumber ← ((weekNumber + i) % 4)
7:         ratingGroup ← ((groupNumber + ratingRotation) % 4) + 1
8:         group ← group with id of ratingGroup
9:         Create a QuestionAnswerer for each member of group
10:    end for
11: end procedure
```

Once questions were assigned, a second, more complicated algorithm was required to assign the previous week's answers to be rated. The pseudocode for this is detailed in Algorithm 2. Much trial and error was spent in the development of this to ensure that a group was never assigned to rate itself. Similarly to Algorithm 1, the current week number must be calculated and saved in the variable *weekNumber*. Additionally, a *ratingRotation* variable is required to hold a value between 0 and 2, which will later be added to the *weekNumber*. In the for loop, similar calculations to the first algorithm are made, along with the additional *ratingRotation* value. Because this algorithm is run a week after the selected questions are answered, the calculation of *groupNumber* without the addition of 1 finds the group number from the previous week. The *ratingGroup* calculation then adds both the additional *ratingRotation* value between 0 and 2, as well as an additional 1. This results in the addition of a value between 1 and 3 to *groupNumber*, which guarantees that the result will never equal the previous *groupNumber* post-modulation. Each group will rate every other group and every question type/category combination 3 or 4 times. It should be noted that because this algorithm requires answers to already be posted, it could not be executed until the second week of the study, and needed to be executed once more after the final week of the study.

Bibliography

[1] [online]URL: http://www.sciencekids.co.nz/quizzes/food.html [cited 1/15/15].

[2] [online]URL: http://www.sciencekids.co.nz/quizzes/technology.html [cited 1/15/15].

[3] [online]URL: http://www.usefultrivia.com/sports_trivia/ [cited 1/15/15].

[4] [online]URL: http://www.knowledgediary.com/article-226.html [cited 1/15/15].

[5] ANSI/NISO z39.87-2006 (r2011) data dictionary [online]. URL: http://www.niso.org/apps/group_public/project/details.php?project_id=69 [cited 10/11/14].

[6] Ask.fm [online]. URL: http://www.ask.fm [cited 2/16/15].

[7] ASP.NET MVC Tutorial [online]. URL: http://www.w3schools.com/aspnet/mvc_intro.asp [cited 2/15/16].

[8] Bing [online]. URL: http://www.bing.com [cited 2/15/16].

[9] Can a video file contain a virus? [online]. URL: https://www.opswat.com/blog/can-video-file-contain-virus [cited 10/29/14].

[10] DMCA - Digital Millenium Copyright Act [online]. URL: https://www.eff.org/issues/dmca [cited 10/29/14].

[11] DMCA Safe Harbor [online]. URL: https://www.chillingeffects.org/topics/14 [cited 10/29/14].

[12] Google [online]. URL: http://www.google.com [cited 2/15/16].

[13] GPS — Nikon [online]. URL: http://www.nikonusa.com/en/Learn-And-Explore/Article/gr9ozfh7/gps.html [cited 10/11/14].

[14] Jelly [online]. URL: http://www.jelly.co [cited 2/15/16].

[15] Jobstr [online]. URL: http://www.jobstr.com/ [cited 2/15/16].

[16] Next generation identification [online]. URL: http://www.fbi.gov/about-us/cjis/fingerprints_biometrics/ngi [cited 10/11/14].

[17] OECD [online]. URL: http://www.oecd.org/unitedstates [cited 2/15/16].

[18] Ruby on Rails [online]. URL: http://rubyonrails.org/ [cited 2/15/16].

[19] StackOverflow [online]. URL: http://www.stackoverflow.com/ [cited 2/15/16].

[20] Standards: Exif [online]. URL: http://www.photometadata.org/meta-resources-metadata-types-standards-exif [cited 10/11/14].

[21] Stutchman Forensic Laboratory [online]. URL: http://stutchmanforensic.com/Services/Voice%20Identification%20Aural%20Comparison%20Collection%20of%20Exemplar.htm [cited 10/7/14].

[22] Techunique apps [online]. URL: http://www.tequnique.com/apps [cited 2/15/16].

[23] TREC - Text Retrieval Conference [online]. URL: http://trec.nist.gov/ [cited 2/16/15].

[24] Voice qualities [online]. URL: http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/quality.html [cited 1/26/15].

[25] Webmasters held personally responsible [online]. URL: http://www.a-w.org/webmaster_held.html [cited 10/29/14].

[26] Yahoo! Answers [online]. URL: http://www.answers.yahoo.com [cited 2/15/16].

[27] Human interaction speeds [online]. August 2000. URL: http://www.humanfactors.com/newsletters/human_interaction_speeds.asp [cited 1/26/15].

[28] FBI launching nationwide facial recognition photo search service [online]. October 2011. URL: https://publicintelligence.net [cited 10/11/14].

[29] *Capitol Records, Inc. v. Thomas-Rasset*, volume 11-2820. 2012.

[30] Just how big are porn sites? [online]. April 2012. URL: http://www.extremetech.com/computing/123929-just-how-big-are-porn-sites/2 [cited 11/7/14].

[31] Dictionary.com unabridged [online]. Oct 2014. URL: http://dictionary.reference.com/browse/voiceprint [cited 2/15/16].

[32] FBI launches a face recognition system [online]. September 2014. URL: http://money.cnn.com/2014/09/16/technology/security/fbi-facial-recognition/ [cited 10/11/14].

[33] Misery merchants. *The Economist*, July 2014.

[34] N.S.A. collecting millions of faces from web images [online]. May 2014. URL: http://www.nytimes.com/2014/06/01/us/nsa-collecting-millions-of-faces-from-web-images.html [cited 10/11/14].

[35] Merriam-webster online dictionary [online]. Jan 2015. URL: http://www.merriam-webster.com/dictionary/multimedia [cited 2/15/16].

[36] B. Abu-Naim and W. Klas. Smart authoring and sharing of multimedia content in personal area networks based on subject of interest. In *Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on*, pages 1–6, July 2014. doi:10.1109/ICMEW.2014.6890603.

[37] Michael Arrington. Yahoo answers launches [online]. December 2005. URL: http://techcrunch.com/2005/12/08/yahoo-answers-launches/ [cited 1/15/15].

[38] Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha. Finding the right facts in the crowd: factoid question answering over social media. In *Proceedings of the 17th international conference on World Wide Web*, pages 467–476. ACM, 2008.

[39] Jiang Bian, Yandong Liu, Ding Zhou, Eugene Agichtein, and Hongyuan 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*, pages 51–60. ACM, 2009.

[40] Frédéric Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, and Douglas A. Reynolds. A tutorial on text-independent speaker verification. *EURASIP J. Appl. Signal Process.*, 2004:430–451, January 2004.

[41] Mohamed Bouguessa, Shengrui Wang, and Benoit Dumoulin. Discovering knowledge-sharing communities in question-answering forums. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(1):3, 2010.

[42] Rebecca Boyle. Biometric voice-identification technology leads to arrests of terrorist suspects in europe [online]. URL: http://www.popsci.com/technology/ [cited 11/7/14].

[43] Bogdan Calin. Why file upload forms are a major security threat [online]. URL: http://www.acunetix.com/websitesecurity/upload-forms-threat/ [cited 11/5/14].

[44] J. Carretero, F. Isaila, A.-M. Kermarrec, F. Taiani, and J.M. Tirado. Geology: Modular georecommendation in gossip-based social networks. In *Distributed Computing Systems (ICDCS), 2012 IEEE 32nd International Conference on*, pages 637–646, June 2012. doi:10.1109/ICDCS.2012.36.

[45] Cisco. Cisco visual networking index: Global mobile data traffic forecast update, 2013–2018 [online]. URL: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html [cited 11/6/14].

[46] Irina Diaconita, Christoph Rensing, and Stephan Tittel. Getting the information you need, when you need it: A context-aware q&a system for collaborative learning. In *Proceedings of the 9th European Conference on Technology Enhanced Learning (Lecture Notes in Computer Science)*, volume 8719, pages 410–415. Springer International Publishing, Sep 2014.

[47] G.R. Doddington. Speaker recognition - identifying people by their voices. In *Proceedings of the IEEE*, volume 73, pages 1651–1664, Nov 1985.

[48] Nick Doty. Strip invisible metadata [online]. URL: `http://privacypatterns.org/patterns/Strip-invisible-metadata` [cited 11/6/14].

[49] F. Maxwell Harper, Daphne Raban, Sheizaf Rafaeli, and Joseph A. Konstan. Predictors of answer quality in online q&a sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 865–874, New York, NY, USA, 2008. ACM.

[50] Lesley Ellen Harris. Legally using images [online]. March 2010. URL: `http://www.copyrightlaws.com/us/legally-using-images/` [cited 10/11/14].

[51] Geoff Herbert. Who owns the pictures you upload on the web? probably not you. [online]. January 2011. URL: `http://www.syracuse.com/have-you-heard/index.ssf/2011/01/who_owns_pictures_you_upload_o.html` [cited 10/11/14].

[52] Debra S. Herrmann. *Complete Guide to Security and Privacy Metrics: Measuring Regulatory Compliance, Operational Resilience, and ROI*. Auerbach Publications, 2007.

[53] Richard Hong, Meng Wang, Guangda Li, Liqiang Nie, Zheng-Jun Zha, and Tat-Seng Chua. Multimedia question answering. *IEEE MultiMedia*, 19(4):72–78, 10 2012.

[54] Grace YoungJoo Jeon, Yong-Mi Kim, and Yan Chen. Re-examining price as a predictor of answer quality in an online q&a site. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 325–328, New York, NY, USA, 2010. ACM.

[55] Dan Kedmey [online]. September 2014. URL: `http://time.com/3246562/hackers-jennifer-lawrence-cloud-data/` [cited 10/11/14].

[56] Balachander Krishnamurthy and Craig E. Wills. On the leakage of personally identifiable information via online social networks. In *Proceedings of the 2Nd ACM Workshop on Online Social Networks*, WOSN '09, pages 7–12, New York, NY, USA, 2009. ACM.

[57] Uichin Lee, Eunhee Yi, and Minsam Ko. Mobile q&a: Beyond text-only q&a and privacy concerns. *In Proceedings of CHI 2013, ACM*, 2013.

[58] Baichuan Li, Tan Jin, Michael R Lyu, Irwin King, and Barley Mak. Analyzing and predicting question quality in community question answering services. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 775–782. ACM, 2012.

[59] Guangda Li, Zhaoyan Ming, Haojie Li, and Tat-Seng Chua. Video reference: question answering on youtube. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 773–776. ACM, 2009.

[60] Ming Li, Shucheng Yu, Ning Cao, and Wenjing Lou. Privacy-preserving distributed profile matching in proximity-based mobile social networks. *Wireless Communications, IEEE Transactions on*, 12(5):2024–2033, May 2013. doi:10.1109/TWC.2013.032513.120149.

[61] Ze Li, Haiying Shen, Guoxin Liu, and Jin Li. Sos: A distributed mobile q&a system based on social networks. In *Proceedings of 2012 IEEE 32nd International Conference on Distributed Computing Systems (ICDCS)*, pages 627–636. IEEE, 2012.

[62] Margaritis Margaritidis and George C. Polyzos. Adaptation techniques for ubiquitous internet multimedia. *Wireless Communications and Mobile Computing*, 1(2):141–163, 2001.

[63] Niall McCarthy. Most social networks are now mobile first [online]. April 2014. URL: http://www.statista.com/chart/2109/time-spent-on-social-networks-by-platform/ [cited 10/7/14].

[64] Mary Meeker. Mary meeker's stunning 2014 presentation on the state of the web [online]. URL: http://www.businessinsider.com/mary-meekers-2014-internet-presentation-2014-5#-62 [cited 2/16/15].

[65] Tao Mei, Yong Rui, Shipeng Li, and Qi Tian. Multimedia search reranking: A literature survey. *ACM Computing Surveys (CSUR)*, 46(3):38, 2014.

[66] Anthony Misquitta. User generated content – risks, prevention and protection for service providers in the uk: Part 1 [online]. URL: http://www.farrer.co.uk/Global/Briefings/16.%20Briefing/User%20Generated%20Content%20pt1.pdf [cited 10/7/14].

[67] Anthony Misquitta. User generated content – risks, prevention and protection for service providers in the uk: Part 2 [online]. URL: http://www.farrer.co.uk/Global/Briefings/16.%20Briefing/User%20Generated%20Content%20pt2.pdf [cited 10/7/14].

[68] Lisa Myers. An exploration of voice biometrics an exploration of voice biometrics an exploration of voice biometrics an exploration of voice biometrics [online]. URL: http://www.sans.org/reading-room/whitepapers/authentication/exploration-voice-biometrics-1436 [cited 10/11/14].

[69] Frank Nagle and Lisa Singh. Can friends be trusted? exploring privacy in online social networks. In *Proceedings of the International Conference on Advances in Social Network Analysis and Mining, 2009. ASONAM'09.*, pages 312–315. IEEE, 2009.

[70] Liqiang Nie, Meng Wang, Zhengjun Zha, Guangda Li, and Tat-Seng Chua. Multimedia answering: Enriching text qa with media information. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 695–704, New York, NY, USA, 2011. ACM.

[71] James O'Toole. Facebook's new face recognition knows you from the side [online]. April 2014. URL: http://money.cnn.com/2014/04/04/technology/innovation/facebook-facial-recognition/ [cited 10/7/14].

[72] Chanmin Park, Seunghyun Kim, and Taehyung Wang. Multimedia copyright protection on the web - issues and suggestions. In *Multimedia (ISM), 2012 IEEE International Symposium on*, pages 274–277, Dec 2012.

[73] Bill Powell. Images in html: Security risk [online]. URL: http://cms.about.com/od/maintain-your-cms-website/a/Images-In-Html-Security.htm [cited 10/29/14].

[74] Mary Rasenberger and Christine Pepe. Copyright enforcement and online file hosting services: Have courts struck the proper balance. *Copyright Society of the USA*, 59:627–695, 2011-2012.

[75] Pew Research. Part 1: How the internet has woven itself into american life [online]. URL: http://www.pewinternet.org/2014/02/27/part-1-how-the-internet-has-woven-itself-into-american-life/ [cited 1/12/14].

[76] Felix Richter. United states top music piracy ranking [online]. URL: http://www.statista.com/ [cited 11/7/14].

[77] Jason Row. This is what you need to know about metadata [online]. November 2013. URL: http://www.lightstalking.com/what-is-metadata/ [cited 10/7/14].

[78] Bill Schiffner. Could you fall victim to crime simply by geotagging location info to your photos? [online]. URL: http://www.digitaltrends.com/photography/ [cited 10/7/14].

[79] Chirag Shah, Sanghee Oh, and Jung Sun Oh. Research agenda for social qanda. *Library and Information Science Research*, 31(4):205 – 209, 2009.

[80] H. Shen, Z. Li, J. Liu, and J.E. Grant. Knowledge sharing in the online social network of yahoo! answers and its implications. *Computers, IEEE Transactions on*, PP(99):1–1, 2014.

[81] Anna Shtok, Gideon Dror, Yoelle Maarek, and Idan Szpektor. Learning from the past: answering new questions with past answers. In *Proceedings of the 21st international conference on World Wide Web*, pages 759–768. ACM, 2012.

[82] History.com Staff. The invention of the internet [online]. 2010. URL: http://www.history.com/topics/inventions/invention-of-the-internet [cited 10/7/14].

[83] Maggy Anastasia Suryanto, Ee Peng Lim, Aixin Sun, and Roger HL Chiang. Quality-aware collaborative question answering: methods and evaluation. In *Proceedings of the second ACM international conference on web search and data mining*, pages 142–151. ACM, 2009.

[84] Jaime Teevan, Meredith Ringel Morris, and Katrina Panovich. Factors affecting response quantity, quality, and speed for questions asked via social network status messages. In *Proceedings of ICWSM*, 2011.

[85] Ellen M Voorhees et al. The trec-8 question answering track report. *TREC*, 99:77–82, 1999.

[86] P.K. Yadav and S. Rizvi. An exhaustive study on data mining techniques in mining of multimedia database. In *Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on*, pages 541–545, Feb 2014.

[87] Liu Yang, Minghui Qiu, Swapna Gottipati, Feida Zhu, Jing Jiang, Huiping Sun, and Zhong Chen. Cqarank: jointly model topics and expertise in community question answering. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 99–108. ACM, 2013.

[88] Lan Zhang, Xiang-Yang Li, Yunhao Liu, QiuYuan Huang, and Shaojie Tang. Mechanism design for finding experts using locally constructed social referral web. In *INFOCOM, 2012 Proceedings IEEE*, pages 2896–2900, March 2012.