The Use of Information Sources by Internet Users in Answering Questions

Sanghee Oh (corresponding author)

School of Information and Library Science, University of North Carolina at Chapel Hill, CB# 3360, 100 Manning Hall, Chapel Hill, NC 27599-3360. shoh@email.unc.edu.

Jung Sun Oh

School of Information and Library Science, University of North Carolina at Chapel Hill, CB# 3360, 100 Manning Hall, Chapel Hill, NC 27599-3360. ohjs@email.unc.edu.

Chirag Shah

School of Information and Library Science, University of North Carolina at Chapel Hill, CB# 3360, 100 Manning Hall, Chapel Hill, NC 27599-3360. chirag@unc.edu.

Abstract

The purpose of this study was to investigate what kinds of sources people prefer to use when they answer questions online, especially, in the context of social Q&A. Social Q&A is a Web-based service, that allows people to ask questions and receive answers from their fellow users. In social Q&A, people often cite sources of information when they answer questions. It could be a name, a short description, or hyperlinks to the original sources. Yahoo! Answers was chosen for this study due to its popularity as a top ranked social Q&A service as well as its capability for separately indicating sources for the answers in its format. We collected data with a crawler that used Yahoo! Answers APIs. A total number of 5,391 sources were identified and analyzed with the following three approaches: (1) source distribution by online accessibility, (2) source distribution by genre, and (3) source distribution by subjects. At the early stage of this study, it was expected that the results of source preferences heavily relied on sources online, since people ask and answer questions on the Web-based service. Nevertheless, it was found that human (56.4%) was the most frequently cited type of source, and it was followed by online (40%) and offline sources (4%). According to the source distribution by genre, human (56.4%) was followed by the Internet (38.1%), books (3.6%), and mass media (1.6%), and the sub-categories of these sources were analyzed. Additionally, the patterns of source distribution were shown differently across subjects. The categories of Health, Home & Electronics, and Society & Culture relied heavily on human sources, while Computers & the Internet included most of the Internet-based sources of information.

Introduction

The use of information sources is one of the most heavily studied areas in information seeking behaviors. The patterns of information seeking behaviors can be inferred by the interactions between subjects and sources of information. Information source gathering and selection criteria in various domains have been investigated extensively, e.g., academic research (Rice & Tarin, 1993; Flaxbart, 2001; Hallmark, 2001; Talja, 2002; Fidel & Green, 2004; Yitzhaki & Hammershlag, 2004), everyday information seeking (Chen & Hennon, 1982; Hektor, 2003; Kari & Savolainen 2003), health information seeking (Bates et al., 2006; Warner & Procaccino, 2004, Case et al., 2004). In addition, the issues related to source accessibility (Fidel, et al., 2004; Xu, Tan, & Yang, 2005), source credibility/authority (Bates et al., 2006), source quality (Choo, Deltor & Turnbull, 2000; Rieh, 2002), and source preferences (Rees & Bath, 2000; Krikelas, 1983) have been widely reviewed and studied in the field.

In the study reported here, we were mainly interested in the source preferences of Internet users in the context of social Questioning and Answering (Q&A). Social Q&A offers a new environment for online users to seek and share information. It is a community-based service, which allows people to ask
questions and receive answers from their fellow users online. People primarily use the service to look for information for solving problems in real life. The topics vary from personal problems to school projects or business problems. People obtain information from the answers, dynamically created and customized, that respond to their original questions.

In the U.S., the popularity of social Q&A services has dramatically increased in recent years. According to the Hitwise report, there was an 889 percent increase in the U.S visits to social Q&A services between 2006 and 2008 (Tatham, 2008). In spite of the increased popularity of the service, there have been few studies that observe and analyze the phenomenon. As one of the early attempts to understand the information seeking behaviors of social Q&A users, this study focused on investigating the information sources of the users, and highlighting the uniqueness of the context based on the findings.

The basic principle of social Q&A is simple: people post questions and their fellow users provide answers to the questions. From the observation of the answers, it was found that answerers often cite sources of information along with their answers in order to prove the validity of their answers. Source information could be a name, a short description, or a URL of the original source. In this study, it was assumed that the use of information sources in social Q&A could be inferred by the analysis of the source preferences of answerers. Thus, the purpose of this study was to investigate what types of sources people prefer to cite when answering questions in social Q&A. In addition, we observed how the source types are distributed in general, as well as across various topical subjects.

This paper is organized in the following way. First, the literature related to source preferences of Internet users is reviewed. Next, we describe the research methods used to collect and analyze the source information cited in answering questions in a social Q&A service, Yahoo! Answers. Finally, the study results are summarized, and discussed.

Source Preferences

According to Krikelas's (1983) early review of the source preferences literature, people use their own experiences and memories first to assess whether they would be able to solve a problem and then ask for assistance to gather additional information from external sources. Among those external sources, people “strongly preferred” human sources, in particular, those transferred through voice communication in face-to-face contact. As one of the factors that influence people to choose certain sources first over others, Krikelas (1983) pointed out that the ease with which the sources are accessed mattered significantly. Thus, people would first ask their immediate family, friends or colleagues because they are usually nearest to them.

The advent of the Internet, however, has dramatically changed the patterns of source preferences. It has allowed people to easily access an immense number of sources beyond their personal contacts. It was observed that the information sources which had been traditionally popular, such as people, newspapers, magazines, and television, have been gradually replaced by Internet sources (Kaye and Johnson, 2003; Hektor, 2003). Interestingly, Case et al. (2004) found that the Internet was the most preferred source among people looking for information about genetics and diseases and constituted 45% of the first choice sources, followed by medical doctors (18.4%), a public library (14.1%), family members (10.6%), medical sources (8.7%) and mass media sources other than the Internet (1%). Mass media sources and sources labeled friends were identified as less usable sources in this study. This change seems surprising since the subject domain was health-related information seeking which would require gathering critical information to make important decisions in treating diseases. People went to the Internet first, even before asking medical experts.

In the meantime, Savolainen and Kary (2004) investigated how Internet users utilize information for solving problems in their real lives and applied a model called an “information horizon” to illustrate the patterns of information source use. The theoretical framework of an “information horizon” was first introduced by Sonnenwald, Wildemuth and Harmon (2001). They used interview and survey methods for testing the framework in understanding the information seeking behaviors of student groups. The information horizon maps are not only helpful for identifying the preferences for information sources used...
by individuals, but also for measuring the relationships and graphically representing the networks and connections among sources. Savolainen and Kary (2004) adopted the framework, renaming it “information source horizon.” The subjects were asked to place their sources in three zones, (1) most significant sources, (2) intermediate sources, and (3) peripheral sources, and to illustrate the different uses of the information sources in each zone. In contrast to the results shown in the previous study by Case et al. (2004), it was found that people still preferred to use human resources in general, followed by print media materials like newspapers and books, and network sources. Savolainen (2007) examined the source preferences of environmental activists with the modified framework of an “information source horizon,” and the result was almost the same as his previous study.

The results from the studies cited here indicate that there is still debate about whether people would prefer to access Internet sources prior to other sources. In our study, it was first assumed that answerers would rely heavily on Internet sources due to several characteristics of social Q&A. First, social Q&A is a Web-based service. It might be easier for answerers to provide sources readily accessible online, by posting direct links to the sources with their URLs. Also, questioners would prefer to have the Internet sources because they would be able to immediately check the source quality by reviewing the links to the sources. Second, in social Q&A, people look for information in order to answer an inquiry by an anonymous stranger, not their own information needs. Thus, it was assumed that people would spend less time and effort searching for answers, and Web searching could be the easiest way to do this for online users. At the end of the study, however, it was found that the Internet sources were not the most preferred source for social Q&A users. This finding will be discussed in the results section in detail.

Another important aspect of the literature on source preferences is the source categories designed for a number of previous studies in order to analyze user preferences in various contexts. Case et al. (2004) classified sources for genetics information seeking into 15 categories including the Internet or Web, public libraries, doctors, family members, cancer information services, hospital programs, magazines, family and friends, newspapers, local and national TV programming, and so on. Stefl-Mabry (2003) was interested in how people express their satisfaction with the information received and developed a source preference profile with six categories including word of mouth, expert oral advice, Internet, reference books, print news, and radio or TV. Xu, Tan, and Yang (2005) focused on evaluating the cost-benefit aspects of sources based on source quality and accessibility. They classified information sources on two dimensions, (1) personal or impersonal, and (2) internal, or external and grouped them into four categories with a combination of these dimensions. Savolainen (2007) used six categories of sources for testing the “information source horizon” framework, including human sources, broadcast media, print media, networked sources, organizational sources, and other sources.

In the literature, the source categories varied according to subject domain and the context of the studies; however, there are three categories that commonly appeared in each study. These are human sources, mass media sources, and Internet sources. Thus, for our study, we started by classifying sources within these three categories. In the process of the analysis, we modified and narrowed them to more specific categories. As a result, a source category scheme, which represented the context of social Q&A, was developed.

Methods

Among various social Q&A services, Yahoo! Answers was a natural choice for this study. The primary reason behind this decision was the fact that Yahoo! Answers was ranked at the top of social Q&A services, which dominated 74.05 % of the market share as of March, 2008 (Hitwise.com, 2008). Since Yahoo! Answers release in December 2005, it has been growing incredibly fast. Yahoo! Answers announced that it hosted 25 million users in the U.S. and 135 million users world wide in 2008 (McGee, 2008). Due to its popularity, it was expected that Yahoo! Answers would reveal the nature of social Q&A more clearly than others, as a representative case.

Another reason that we chose Yahoo! Answers for this study was its format, namely, the capability of separately indicating sources for the answers. In the interface of Yahoo! Answers, there is a field to fill out in which answerers fill out the source of the information when posting answers. Yahoo! Answers
encourages answerers to supply source information in the field by providing points to answerers for posting sources. Thus, although it is not required, we found that a good number of answerers use it. In reality, many answerers provide the source information within their answers, integrating it into the contents of their answers, not indicating source information separately in the source field. However, considering that Yahoo! Answers grants additional points for specifying sources in the source field, we assume people who can clearly point to the source of their answers are likely to include such information in the designated field. Therefore, given the huge volume of data, it was decided to collect only those data shown in the source field for the analysis in this study.

In order to collect data for our study, we created a crawler that used Yahoo! Answers APIs (http://developer.yahoo.com/answers/) as well as various screen-scraping techniques. Since Yahoo! has imposed a restriction of up to 5,000 API calls per 24 hours only, there was a limit to how much data we could collect in a given day. Within this framework and restrictions, we collected nearly 250,000 questions with more than 3 million answers during the last three months of the year 2007. Each question and answer was stored with the content of the post, and the subject category.

From this collection, we randomly extracted 13,167 questions for our study. The number of answers linked to those questions was 101,985. Thus, the average number of answers per question in our sample was 7.75. Gazan (2007) conducted a study with AnswerBag about the use of the social Q&A service in finding answers to homework questions of students, and the average number of answers per question was 1.3 in the context of his study. In contrast to this, a much higher number of average answers per question for Yahoo! Answers were observed in our study, and it was encouraging from the user participation point of view. Among the 101,985 answers, 7,834 answers (7.68%) included source information in the source field of the answers. With a semi-automatic coding process, we were able to extract a total of 8,119 units of sources from answers and these became the units of analysis for our study.

The procedure of the source analysis was as follows. From the initial review of the source data, we were able to identify a set of keywords to represent the source information such as domain names of Websites. Thus, we semi-automatically extracted those keywords as information about sources first. At the same time, we filtered out keywords not particularly related to source information. Since Yahoo! Answers provides reward points to the answerers when they write something in the source field, there are some occasions in which the words are not quite related to sources. Thus, we filtered out symbols such as emoticons, or punctuation. Examples included writing an email address for further contact or consultation about questions, or simply words such as “me”; these were excluded during the analysis process. Finally, a total of 5,391 sources were obtained for further analysis.

With the collected source data, three approaches were used to review and categorize the sources. First, the sources were categorized depending on their online accessibility – human, online, and offline. Second, the sources were classified by genre. This classification was based on the basic three types of sources - human, mass media and the Internet - from the previous studies, but the scope of each type was newly defined and sub-levels of categories were specified giving consideration to the unique patterns of sources in social Q&A. Finally, the source distribution was compared across topical subjects. When questioners post questions in Yahoo! Answers, they need to identify the subject category of the questions, and answerers may provide information with sources responding to the questions, naturally considering the subjects. Thus, in this third approach, it was assumed that the source preferences of answerers could vary depending on the topical subject with which they are dealing when answering questions. For the analysis, the source distribution of high-level genre categories - human, mass media & books, and the Internet, - was observed across subjects.

Results

Source Distribution by Online Accessibility

Since Yahoo! Answers is a Web-based service, it was expected that answerers would prefer to cite online sources over human-related information or offline sources. Thus, in this analysis, we focused on whether
sources are available online or not, and investigated the proportions of the three different types of sources – human, online, and offline. Human sources are the ones which answerers present their own knowledge or the knowledge of third-parties as sources of answers. Answerers released information about their knowledge, experiences, occupations, or degrees as sources of their answers. Online sources are the ones with information of URLs or online Websites or services. Offline sources are the ones with information about names or descriptions of offline materials like books or quotations.

Among the total number of 5,391 sources, 3,041 (56.4%) were identified as human-related sources. Despite the fact that the data were collected from an online service, human sources were still the most frequently cited information sources. As for the rest, online sources (2,144 sources, 39.8%) were cited more frequently than offline sources (206 sources, 3.8%). The number of online sources cited was over 10 times the number of offline sources cited. It was obvious that people prefer to use online sources to offline sources.

Source Distribution by Genre

According to the reviews of previous studies of source preferences, it was found that human, mass media, and the Internet are the three most popular genres of information sources, in general. Therefore, we first looked closely at the distribution of our source data within the basic schema of three genres. Then, we specified the three genres based on the scope of sources in the social Q&A context, and developed a new source classification schema.

The high-level of the source distribution by genre was classified into four categories: human, books, mass media, and the Internet. Although books were not considered as a commonly cited source in previous studies, this category was newly added to the genre analysis in this study due to the recognizable number of sources in our context of social Q&A. The genre categories were defined as follows.

The scope of human sources is the same as the one defined in the previous section. Human sources indicate the knowledge, experience, occupations or jobs of the source providers (answerers in the context of our study of Yahoo! Answers) or the respective third-parties. Book sources indicate mostly printed manuscripts or publications, but the links or URLs of online books were also included in this category. The title or the authors of the books were shown as sources. Mass media indicates sources with information about broadcasts, news, magazines, or television. For example, the name of a broadcast program or a website of the same program may be cited as source information. Thus, it includes both online and offline information about mass media. The rest of the online sources which are not included in any genre of human, books, or mass media were grouped into the Internet sources.
The source distribution of the high-level category of human, books, mass media and the Internet was shown in Figure 1. Human (56.4%) was the most frequently cited source, and was followed by the Internet (38.1%), books (3.6%), and mass media (1.6%). In previous studies, mass media was indicated as a relatively high priority source of information (Steff-Mabry, 2003; Savolainen, 2007). It, however, was less frequently cited as the source in the online environment of this study. The sources which would not be included in any one of these categories, (e.g., quotes from poems, lyrics, music CDs or movie names) were classified as others (0.3%).

The sub-level categories of each type have been further analyzed in order to understand the characteristics of each category in detail. The results have been shown in the following tables (Table 1, Table 2, & Table 3).

Table 1. Human Sources

<table>
<thead>
<tr>
<th>Source types</th>
<th>No. of Sources (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Personal / Situational Experiences</td>
<td>1,186 (39.0 %)</td>
</tr>
</tbody>
</table>
First, human sources were divided into five subcategories, as shown in Table 1. Personal/situational experience (39.0%) was the most frequently cited source of information among the human sources. This indicates that the sources were identified through the explanations of source providers’ experiences. For example, responding to the questions asking about what to do in a situation concerning childcare, experiences as mothers or years of being mothers were provided as sources along with answers to the questions. Thus, experiences in marriage, family, death, disease treatment, travel or any situations similar to those of the questioners were explained and used as sources of information. Professional/educational background (33.0%) was the next most frequently cited source. Answerers often provided information about their occupations or what kinds of degrees they had obtained related to the questions, as well as how long they had been working in the area.

Not only did answerers describe themselves as sources, but also described the third parties. Third-party sources (2.6%) include the sources which referred to the speeches or stories of famous people, like Albert Einstein, or professions or experiences of family members, friends or neighbors. As the fourth type of human sources, research/knowledge (1.9%) indicates sources with information about the results of personal research, theories, phenomenon, or rules of science. Finally, ethnicity/origin (1.9%) includes sources of information about the races of the answerers. For example, in the case of questions that asked about the various racial points of view, the answerers provided their ethnic information as sources. The human-related sources which didn't specify any information related to the previous category were grouped with the others (21.5%). A self-indication such as Christian, atheist, peace lover, etc., was included here, too. Simple comments like “personal experience” with no further information were also treated as others.

<table>
<thead>
<tr>
<th>Source types</th>
<th>No. of Sources (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional / Educational Background</td>
<td>1,003 (33.0%)</td>
</tr>
<tr>
<td>Third-party</td>
<td>79 (2.6%)</td>
</tr>
<tr>
<td>Research / Knowledge</td>
<td>61 (2.0%)</td>
</tr>
<tr>
<td>Ethnicity / Origin</td>
<td>57 (1.9%)</td>
</tr>
<tr>
<td>Others</td>
<td>655 (21.5%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3,041 (100%)</strong></td>
</tr>
</tbody>
</table>

Not only did answerers describe themselves as sources, but also described the third parties. Third-party sources (2.6%) include the sources which referred to the speeches or stories of famous people, like Albert Einstein, or professions or experiences of family members, friends or neighbors. As the fourth type of human sources, research/knowledge (1.9%) indicates sources with information about the results of personal research, theories, phenomenon, or rules of science. Finally, ethnicity/origin (1.9%) includes sources of information about the races of the answerers. For example, in the case of questions that asked about the various racial points of view, the answerers provided their ethnic information as sources. The human-related sources which didn’t specify any information related to the previous category were grouped with the others (21.5%). A self-indication such as Christian, atheist, peace lover, etc., was included here, too. Simple comments like “personal experience” with no further information were also treated as others.

Table 2. Mass Media & Book Sources

<table>
<thead>
<tr>
<th>Source types</th>
<th>No. of Sources (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass Media</td>
<td></td>
</tr>
<tr>
<td>TV, News, Magazine, etc.</td>
<td>87 (46.6%)</td>
</tr>
<tr>
<td>Books</td>
<td></td>
</tr>
<tr>
<td>The Bible</td>
<td>130 (22.2%)</td>
</tr>
<tr>
<td>Other Books</td>
<td>62 (31.2%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>279 (100%)</strong></td>
</tr>
</tbody>
</table>

Mass media and book sources were grouped together in Table 2. As explained in the introduction to the source distribution by genre, mass media includes both online and offline media programs including broadcasts, news, and magazines. In the book category, one third of the sources were about the Bible. These ranged from general indications like the words “Bible,” “The Old Testament,” or “The New Testament” to quotations of scriptural phrases. The online versions of the Bible or related Websites were included here as well. One of the possible reasons that the Bible was shown as the most frequently cited book was because a great deal of data we collected included religious questions, answers and sources on the subject of Society & Culture. (The description of the subject distribution of our source data was illustrated in detail in the following section of the source distribution across the subjects.)
The Internet sources were classified into six sources according to the use of Websites. The domain names of URLs of sources such as .com, .net, .gov, .org, and .edu were used to automatically cluster the Internet resources into the four genres of commercial/personal, organizations, government, and education. In addition, the sources were manually reviewed to identify other source genre, and it was observed that there were a number of sources indicating search engines like Google or Yahoo!, and providing links to the results pages of those search engines. These sources with information of search engines were grouped into search engines, separately from the commercial/personal Website group, although their domain names are.com. In addition, there was a pattern of self-citing such as using answers in Yahoo! Answers as the source of information. Those sources were also grouped separately.

As shown in Table 3, a majority of the Internet sources were from the commercial/personal category (71.4%). Numerous and various kinds of Websites in the .com and .net domains were grouped into this category and the average frequency of sources per unique domain was 1.6. Organization Websites (16.9%) was the second biggest category of the Internet sources. Wikipedia belongs to this category. Among the 348 sources of organization Websites, 125 sources (35.9%) were from Wikipedia. The popularity of Wikipedia as an information source can be seen from this result. Search engines (4.4%) were also considered by people to be sources, and people cite the answers from Yahoo! Answers (59%) as sources. The use of other social Q&A services such as answers.com or AnswerBag was negligible. In order to closely look at the Website preferences cited as sources, the Websites, most frequently cited as sources are listed in the discussion section of this paper.

### Source Distribution across Subjects

When questioners post questions in Yahoo! Answers, they need to identify the subject category of the questions, and answerers may provide information with sources responding to the questions, naturally considering the subjects. Thus, in this third approach, it was assumed that the source preferences of answerers could vary depending on the topical subject with which they are dealing when answering questions. For the analysis, the source distribution of high-level genre categories - human, mass media & books, and the Internet, - was observed across subjects.

Yahoo! Answers originally offered 25 subject categories. In order to present the distribution concisely, 25 subject categories were regrouped into 12, based on the comparison with the Internet subject category schema offered by Librarian’s Internet Index (LII, [http://lii.org](http://lii.org)) and the Internet Public Library (IPL, [http://www.ipl.org](http://www.ipl.org)).

In our data set of sources, Society & Culture (34.2%) was the most frequently shown subject followed by Recreation & Hobbies (27.9%), Health (8.6%), Politics & Government (6.9%), Home & Electronics (5.9%), Business & Finance (4.5%), Computers & Internet (3.3%), Education & Reference (2.7%), Arts & Humanities (2.5%), Science & Mathematics (1.8%), Social Science (1.0%) and News & Media (0.6%). In order to compare the ratio of human, mass media & books, and the Internet sources in each category, the frequency data of sources in each category was transformed to a percentile graph and illustrated in Figure 2.
The categories of Health, Home & Electronics, and Society & Culture relied heavily on human sources. In the Health category, half of the human sources belonged to the sub-category of professional/educational background (55.2%), such as describing medical-related degrees or number of years as a registered nurse. The personal/situational experiences (23.8%) category was followed with examples of people’s experience with similar diseases or health care critiques. In the Home & Electronic category, the professional/educational background (58.0%) and personal/situational experiences (27.4%) were also the most popular categories within human sources. Careers as engineers, designers, or photographers and experiences using particular appliances or devices were recognized as sources in these categories. The pattern of the sub-categories of human sources in the Society & Culture category is a little different from the previous two subject categories. Since the childcare related information was classified in this category, the sources related to personal/situational experiences (50.4%), such as years of being a mother and the number of pregnancies were often described.

On the other hand, as shown in Figure 2, Computers & the Internet included the largest proportion of Internet-based sources of information. Among the Internet sources, the sub categories of commercial/personal Websites (78.2%) were cited most frequently. In a majority of the cases, people posted URLs of Websites or specific names of Websites as sources in order to confirm where they copied...
the content of answers responding to questions asking about problems with computer software or hardware.

Mass media & book sources were cited most frequently in the News & Events category. Society & Culture was the subject category with a relatively large proportion of mass media and book sources since it included the subject category of religion, and many answers were proposed with the Bible as the source. In the rest of the subject categories, human sources and the Internet sources were shown almost equally frequently.

Discussion

This study was first designed with the assumption that Internet sources would be the most frequently cited by answerers, but it turned out that human sources were still the top sources of information in social Q&A. Strictly speaking, the origin of human sources in the context of social Q&A is a little bit different from that in previous studies in terms of who provides the information. In the previous studies, human sources were often indicated as the people who are familiar with and physically available to information seekers, such as their family members, friends, or neighbors (Krikelas, 1983; Case et al., 2004). In social Q&A, people obtained information (answers) from their fellow users, who are anonymous in the online environment. In spite of the lack of familiarity, information seekers still preferred to obtain information from other human beings. In terms of descriptions of human sources in social Q&A, answerers described their knowledge, experience, or occupations and emphasized that they were credible sources of answers by highlighting their expertise or experiences and by articulating how many years or how long they had been professional or in relevant situations (e.g., “14+ years experience as a paralegal specialist,” “French teacher since 1978,” “Vegetarian for over 30 years”).

Internet sources were the second most cited sources in social Q&A. The links to Websites or specific pages or portions of Web pages were attached as sources corresponding to the answers with the content. The estimation of the frequencies of the unique domains of Websites presented interesting results. After removing the duplicates among 2,144 Internet sources, a total of 1,308 unique domains were identified, and the average frequency of each Website cited as a source was 1.6 times. The frequencies of the top 60 Websites occupied one third of the total number of frequencies. Table 4 showed the top 10 Websites frequently cited as sources.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Websites</th>
<th>No. of Sources (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Wikipedia</td>
<td>126 (5.9%)</td>
</tr>
<tr>
<td>2</td>
<td>Yahoo! Answers</td>
<td>61 (2.8%)</td>
</tr>
<tr>
<td>3</td>
<td>YouTube</td>
<td>53 (2.5%)</td>
</tr>
<tr>
<td>4</td>
<td>Google</td>
<td>46 (2.1%)</td>
</tr>
<tr>
<td>5</td>
<td>Yahoo! Search</td>
<td>44 (2.1%)</td>
</tr>
<tr>
<td>6</td>
<td>Blogspot.com</td>
<td>36 (1.7%)</td>
</tr>
<tr>
<td>7</td>
<td>IMDB</td>
<td>25 (1.2%)</td>
</tr>
<tr>
<td>8</td>
<td>About.com</td>
<td>20 (0.9%)</td>
</tr>
<tr>
<td>9</td>
<td>Amazon.com*</td>
<td>14 (0.7%)</td>
</tr>
<tr>
<td>9</td>
<td>WebMed*</td>
<td>14 (0.7%)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>2,144 (100.0%)</td>
</tr>
</tbody>
</table>

*Amazon.com and WebMed were tied and ranked 9th together; there is no source ranked 10th.

As can be seen in Table 4, Wikipedia was the most frequently cited Website. According to the subject category distribution of sources, Wikipedia was most often cited in the answers in the categories of
Society & Culture (32.5%) and Recreation & Hobbies (24.6%). It was also often cited as a source in the areas of Science & Mathematics (14.3%), Arts & Humanities (5.6%), Politics & Government (7.8%) and Education & References (4.8%). It seems that the influence of Wikipedia was distributed across the subjects and used as a common source of information.

The second ranked Website was Yahoo! Answers. People referred to the content of answers posted in Yahoo! Answers as the source of their other answers. Under this kind of cycle of self-citing, an answer was reused in another answer. Furthermore, since answers are publicly available and retrievable, they could have been recycled for various purposes of information distribution beyond their original function as a response to a particular question. Currently, the answers of Yahoo! Answers are crawled by search engines such as Google, and displayed in the search result pages (sometimes near the top of the results list). This will lead more and more people to access and reuse the answers.

When people used Google or Yahoo! Search as sources, some of them included the specific URLs of the search result pages, whereas some of them literally mentioned Google or Yahoo! Search as sources without additional information about the section or services to which they were referring. Blogspot.com was ranked as the sixth most cited Internet source. Other blogs were also cited frequently as sources in answers. The popular Internet Websites, such as YouTube, IMDB, About.com, Amazon.com, and WebMed, were among other popular sources cited in answers.

Limitations of the Study & Future Work

There are two limitations that need to be acknowledged and addressed regarding the current study. First, in data collection, using the source information extracted only from a designed field of Yahoo! Answers saved time and energy collecting and analyzing the data. It caused, however, an isolation of the source information from questions and answers, and the exclusion of the source information that is integrated within answers from the analysis. Second, the descriptive statistics of the analysis, based on the counts of the frequencies of source citation, contributed to illustrate the big picture about the use of information sources, but it was insufficient to explain the dynamics of the source selection and citation behaviors of users corresponding to various situations of information seeking and sharing in social Q&A. Considering these limitations, in future study, the use of the source information will be examined in relation to questions, as well as answers and user situations, and the in-depth analysis to identify the uncovered diversity of user behaviors the will be conducted.

Conclusion

The current investigation centered on how the source preferences of information seekers have changed in the online environment, in particular, in the context of Social Q&A. Interestingly, human sources were found to be the most frequent sources of information, just as in the context of traditional information seeking environments. People noted their various backgrounds and prior knowledge as the sources of their answers, as well as their personal and situational experiences. The use of online sources dramatically increased in the Social Q&A context, while the use of mass media or books as information sources decreased when compared to the data shown in other studies. It was also found that the source preferences varied across the subject areas of the questions.

Note


References


