8-2010

Answer Reliability on Q&A Sites

Pnina Shachaf
Indiana University, shachaf@indiana.edu

Answer Reliability on Q&A Sites

Pnina Shachaf
Indiana University, Bloomington
shachaf@indiana.edu

ABSTRACT

Similar to other Web 2.0 platforms, user-created content on question answering (Q&A) sites raises concerns about information quality. However, it is possible that some of these sites provide accurate information while others do not. This paper evaluates and compares answer reliability on four Q&A sites. Content analysis of 1,522 transactions from Yahoo! Answers, Wiki Answers, Askville, and the Wikipedia Reference Desk, reveals significant differences in answer quality among these sites. The most popular Q&A site (that attracts the largest numbers of users, questions, and answers) provides the least accurate, complete, and verifiable information.

Keywords

Askville, WikiAnswers, Wikipedia Reference Desk, Yahoo! Answers, Q&A sites, CQA, social reference, web 2.0, quality, crowd sourcing.

INTRODUCTION

IS researchers are increasingly paying attention to question answering (Q&A) sites (Ong, Day & Hsu, 2009); yet, research on Q&A sites is still in its infancy, partially due to the newness of the phenomenon. At the same time, visits to Q&A sites have increased tremendously over the last few years (Hitwise, 2008). Online communities of volunteers on Q&A sites are processing millions of questions and answers online. Yahoo! Answers is among the most frequently consulted reference sites, second only to Wikipedia. It has over 100 million users and more than 23 million archived questions (Adamic et al., 2008). Q&A sites, like other Web 2.0 social sites, take advantage of the wisdom of the crowds (Surowiecki, 2004), and rely on users' participation (O'Reilly, 2005); they exemplify the idea of social reference by crowd-sourcing question answering work (Shachaf, 2009). As part of the social reference, the roles of information professionals and consumers blur to facilitate growing profits. It is critical to evaluate the quality of the outputs produced by means of crowd-sourcing. The possible benefits of cost reduction and user participation are attractive features of crowd-sourcing, yet a major risk includes the potential provision of inferior services.

Researchers from various disciplines (information retrieval, human computer interaction, reference, information seeking behavior and use) are trying to grasp the depth and range of impact that Q&A sites have on traditional conceptions of information creation, dissemination, intermediation, and use (Shachaf & Rosenbaum, 2009). Answer quality is a common concern among all of these disciplines. The social Web, and Q&A sites in particular, forces researchers to raise questions about the reliability of user-created content. Research provides evidence that the quality of Wikipedia is as good as traditional encyclopedias (e.g., Giles, 2005). Q&A sites, like Wikipedia, build on the idea that everyone knows something and through collaborative knowledge production users provide answers to questions and create an archive of questions and answers. The costs and benefits of these sites are questionable. Do they pose a threat to our culture and traditional institutions by supporting a culture of mediocrity where *everything is miscellaneous* (Wienberger, 2007) and by fostering a *cult of amateurs* (Keen, 2008)? Or do they provide an opportunity for further development of our traditional institutions and practices by providing high quality information to users? More specifically, the following questions should be addressed: What is the quality of services (answers) that users receive on these Q&A sites? Are these answers accurate, reliable, and complete? Do some of these sites provide better services than others? One of the first steps in answering these questions relies on a rigorous quality assessment of these sites. This paper focuses on this lacuna and examines and compares answer quality on four Q&A sites: Askville, WikiAnswers, the Wikipedia Reference Desk, and Yahoo! Answers through on an analysis of 1,522 randomly selected transactions. [1]

[1] A transaction refers to a question and all of its respective answers.

BACKGROUND

As the popularity of Q&A sites increases among users, scholarly interest in answer quality increases as well. Researchers have mostly focused on Yahoo! Answers and have based their assessment on user rankings of best answers (e.g., Adamic et al., 2008). A few studies of answer quality extend beyond Yahoo! Answers (Harper et al., 2008; O'Neill, 2007; Shachaf, 2009). Harper at al. (2008) found that answer quality on Google Answers is better than Yahoo! Answers, AllExperts, and libraries. Shachaf (2009) found that answers at the Wikipedia Reference Desk are as good as answers that librarians provide. Yet, O'Neill (2007, p. 10) argues that "responders at Yahoo! Answers and Askville could find it difficult to handle questions that really require an old fashioned reference interview and/or some knowledge of resources not easily uncovered by simple search." The utility of these studies is limited because they either focus mostly on services that seize operations (for example, Harper et al., 2008).[2] Others (for example, O'Neill, 2007) rely on a small sample of questions from Yahoo! Answers, ChaCha, and Askville, or focus on sites that are not the most typical of Q&A sites (for example, Shachaf, 2009), such as the Wikipedia Reference Desk. These earlier studies suggest that answer quality varies across Q&A sites, but it is unclear which of the existing Q&A sites is better than the others. This paper aims to address this lacuna by conducting a comparative analysis of some of the existing and most popular Q&A sites, focusing attention on answers reliability.

Measuring answer reliability, as an indication of answer quality, is not as common in Q&A research as "best answers" or user reputation, yet answer reliability can serve well to examine answer quality across sites. Answer reliability can be determined through an analysis of the content of answers. These type of answer quality measures have been utilized by scholars to evaluate answer quality on Q&A sites, continuing a tradition of assessment of answer quality provided by information professionals (e.g., Harper et al., 2008; Shachaf, 2009). Building on this tradition of references research, Harper et al. (2008) and Shachaf (2009) examined answer quality on Q&A sites. Others have proposed frameworks for Q&A sites that are based on reference research and focused on answer characteristics such as accuracy and completeness of the answer (Blooma, Chua, & Goh, 2009; Ong, Day, & Hsu, 2009). These measures are more objective than "best answers", user reputation, or user satisfaction. Under this approach, high quality answers were determined, for example, based on content analysis (Blooma, Chua, & Goh, 2009; Gazan, 2006; Harper et al. 2008; Shachaf, 2009). Other scholars that analyzed the content of answers have found that better answers are longer (Adamic et al., 2008; Blooma, Chua, & Goh, 2009; Harper et al. 2008) or include references to external sources (Gazan, 2006). Interestingly, Blooma, Chua, and Goh, (2009), report that question category, answer accuracy and completeness, and length of answer are significant predictors of answer quality, whereas asker's and answerer's authority and reputation are not.

Thus, the present study utilizes three reliability measures (Blooma, Chua, & Goh, 2008; Ong, Day, & Hsu, 2009; Shachaf, 2009) that are not confined by site-specific criteria: accuracy, completeness, and verifiability. These three quality measures are used next to compare several Q&A sites and to test whether there are variations in quality across Q&A sites; specifically, the study tests the following hypotheses:

H1: Answer accuracy will vary across Q&A sites.

H2: Answer completeness will vary across Q&A sites.

H3: Answer verifiability will vary across Q&A sites.

Further, because Shachaf (2009) suggested that the quality of an individual answer is inferior to the quality of an amalgamated response (an amalgamated response is composed of multiple answers posted by multiple users answering one question), the study further hypothesizes that:

H4: Quality of an amalgamated response will be better than quality of "best answer."[3]

---

[2] The implications of this study are limited because Google Answers seized operation in 2006, Microsoft shut down its QnA site in 2009, AllExpert does not support collaborative question answering (because only a few members can answer and the question and answer are posted only after an answer is provided), and libraries significantly differ from all these other CQA (Harper et al., 2008; Shachaf, 2010).

[3] "Best answer" is an answer that has been chosen as the best among all answers in an amalgamated response. "Best answer" is common in sites such as Yahoo! Answers, Answerbag, and Askville, but does not exist in site such as WikiAnswers and the Wikipedia Reference Desk. "Best answers" are determined based on site-specific criteria by the user who ask the question or by a community vote.

METHOD

Data collection and analysis focused on four Q&A sites: Askville, WikiAnswers, the Wikipedia Reference Desk, and Yahoo! Answers. Askville, an Amazon Q&A community, was created by Joseph Park, Fai Leong, and Christian Cabanero and was launched in December 2006. WikiAnswers was founded by Chris Whitten in 2002 as FAQ Farm and was acquired by Answers Corporation (Answers.com) in November 2006. The Wikipedia Reference Desk was launched in 2001 as part of the Wikipedia project but the reference desk was not heavily used during the first couple of months. Larry Sanger (co-founder of Wikipedia and founder of Citizendium) asked the first question. Yahoo! Answers was launched in December 2005 by Yahoo Inc. and become the most popular Q&A site soon after (Hitwise, 2008).

Data Collection

A random sample of 1,522 transactions from these four Q&A sites was collected. The sample includes transactions from Yahoo! Answers (N=584), WikiAnswers (N=605), Wikipedia Reference Desk (N=77), and Askville (N=256). First, for a pilot study, data from the Wikipedia Reference Desk was collected and analyzed; these transactions included questions and answers posted on April 2007 on all seven topical categories (Shachaf, 2009). Next, for the follow up study, data was collected from Yahoo! Answers, WikiAnswers, and Askville. Three Perl programs were used to harvest transactions from each of these three Q&A sites. The programs were set up to collect the most recent questions per category over a 24-hour period at a random minute of every hour and to collect all the relevant answers that were posted over the following 18 days.[4]

Table 1. Q&A Sites Response Rates

Q&A Site	**Response Rate**	
	Answers/Questions	Percent
Askville	189/256	74%
WikiAnswers	99/605	16%
Wikipedia Reference Desk	74/77	96%
Yahoo! Answers	497/584	85%

The response rates on each of these sites vary (Table 1). The highest response rate was found on the Wikipedia Reference Desk (96% of the questions are answered) and the lowest response rate was found on WikiAnswers (16% of the questions are answered). Given that WikiAnswers is the second most popular Q&A site (Hitwise, 2008), the low response rate is surprising.

In order to verify that the WikiAnswer data set is not corrupted, additional examination of randomly selected 50 unanswered questions was conducted a year after data collection. Using the history feature on each of these questions, an effort was made to determine if there was no answer on the day of data collection. Nearly all of the 50 questions either haven't been answered or were answered (33/50) after the date of retrieval (14/50), and then there were only a few that were merged or were called the same as another question (3/50). Next, to understand if there are similarities and differences between these questions that are answered and those that are not, a random sample of 100 questions from the entire data set was examined. The vast majority of the questions (83 questions) were added to a category, but only 25 questions were answered (four were answered by the date of data collection), additional 26 questions were either confusing or made no sense, but there was no clear reason for the other 49 questions to remain unanswered.

Data Analysis

Content analysis of each of the 1,522 transactions was conducted to identify reliable answers. "Content analysis is an empirically grounded method, exploratory in process and predictive or inferential in intent." (Krippendorff, 2004, p. xvii) Researchers have frequently used content analysis as a method to evaluate answer quality on Q&A sites (e.g., Blooma, Chua, & Goh, 2009; Gazan, 2006; Harper et al. 2008; Shachaf, 2009). Content analysis of answers enables the evaluation of

[4] The sample includes question that have been posted on Sept 30, 2008 and all of their respective answers, which were posted by October 18, 2008.

answer quality based on quantifying the presence or absence of quality codes in the answer text; it is used here to compare answer quality across multiple Q&A sites.

Content analysis was done at the transaction level, focusing on answer quality and using the following three codes to indicate answer reliability:

1. Accuracy of an answer refers to a correct response.
2. Completeness of an answer refers to an answer that is thorough, provides enough information, and answers all parts of a multi-parts question.
3. Verifiability of an answer refers to a response that provides a link or a reference to another source where the information can be found.

First, frequency tables were created for each Q&A site tallying the presence of codes for all answers from a specific Q&A site. Then, a comparative table was created where percentages of codes per Q&A site were marked. Finally, based on these comparative tables, statistical analysis using SPSS 17.0 was done to examine if the differences were statistically significant.

FINDINGS

On each of these four sites, the level of accuracy, completeness, and verifiability was assessed at the transaction level (N=1522). In addition, reliability of the best answer was assessed for all the transactions from two of the four Q&A sites, Askville and Yahoo! Answers (N=1,356).[5]

Table 2. Answer Reliability

		Askville	**Yahoo! Answers**	**Wikipedia**	**WikiAnswers**
Amalgamated Response	Accuracy	47%	32%	56%	53%
	Complete	84%	75%	63%	77%
	Verifiable	43%	25%	76%	6%
Best Answer	Accuracy	50%	37%	NA	NA
	Complete	82%	60%	NA	NA
	Verifiable	39%	9%	NA	NA

In order to examine the first three hypotheses, the level of each of the three reliability measures - accuracy, completeness, and verifiability – was examined for each of the four sites and cross-tabulation analysis was then conducted. The four Q&A sites differ on all three reliability measures (Table 2), but these differences are statistically significant only for two of the measures (Table 3). Cross-tabulation results show that the difference in the level of accuracy between the four sites is statistically significant, $\chi^2(1, N=400)=3.25$, $p=.07$. The Wikipedia Reference Desk provides the most accurate level of answers among the four Q&A sites. The first hypothesis (H1) was therefore supported. Completeness level varies across sites but cross-tabulation results indicate that these variations are not statistically significant, $\chi^2(1, N=400)=.10$, $p=.74$. Thus, the second hypothesis (H2) was not supported. These sites differ significantly in the level of answer verifiability, $\chi^2(1, N=400)=74.82$, $p=.00$, where the Wikipedia Reference Desk provides the most verifiable answers. The third hypothesis (H3) was supported.

[5] "Best answers" are not selected on the WikiAnswers or the Wikipedia Reference Desk.

Table 3. Answer Reliability Across Four Sites

Variable (N=400, df=1)	χ^2	**Cramer's V**	***p* level**
Accuracy*	3.25	.09	.07
Complete**	.10	.01	.74
Verifiable***	74.82	.42	.00

*sig. < .1 ** sig < .05 ***sig. < .01

To examine the fourth hypothesis, the level of each of the three reliability measures was assessed for the best answer on data from the two sites, which have an indication of "best answer", Askville and Yahoo! Answers (Table 2). Cross-tabulation analysis was conducted to examine if the quality of the best answer differs from the quality of an amalgamated response. Results indicate significant differences for two of the three reliability measures on Yahoo! Answers (Table 4). Completeness and verifiability levels were significantly better for amalgamated responses compared with best answers on Yahoo! Answers (for completeness $\chi^2(1, N=200)=5.12$, $p=.02$; and for Verifiability $\chi^2(1, N=200)=9.07$, $p=.00$); this trend was observed on Askville as well, but the differences there were not statistically significant (for completeness $\chi^2(1, N=200)=1$, $p=.31$; and for Verifiability $\chi^2(1, N=200)=.33$, $p=.56$). It is important to note, however, that accuracy levels did not significantly differ between best answers and the amalgamated responses on Yahoo! Answers $\chi^2(1, N=200)=.55$, $p=.45$, and did not differ significantly on Askville $\chi^2(1, N=200)=1.8$, $p=.67$. Interestingly, the trend was opposite for accuracy compared with the other two measures; best answers were more accurate than the amalgamated response on both of these Q&A sites. Thus, the findings partially support the fourth hypothesis (H4); an amalgamated answer on Yahoo! Answers is better than "best answer" in terms of completeness and verifiability.

Table 4. Answer Reliability of Best Answer vs. Amalgamated Response

	Variable (N=200, df=1)	χ^2	**Cramer's V**	***p* level**
Askville	Accuracy	1.8	.03	.67
	Complete	1	.06	.31
	Verifiable	.33	.04	.56
Yahoo! Answers	Accuracy	.55	.05	.45
	Complete**	5.12	.16	.02
	Verifiable***	9.07	.21	.00

* sig. < .1 **sig. < .05 ***sig. < .01

The findings support the argument that answer reliability varies on Q&A sites, and provide evidence that the Wikipedia Reference Desk answers are more accurate and verifiable than the other three more popular Q&A sites, Yahoo! Answers, WikiAnswers, and Askville. The most popular Q&A site, Yahoo! Answers, is the least accurate and, along with WikiAnswers, provides answers with the lowest level of verifiable information. The findings also support the argument that an amalgamated response is better than the "best answer". On Yahoo! Answers an amalgamated response was significantly more complete and verifiable, yet not more accurate than the "best answer".

DISCUSSION

Two major questions surfaced from the findings and will be discussed next: 1) What are some of the reasons for the significant variations in quality of answers across Q&A sites? 2) How can the mixed findings about the quality differences between "best answers" and amalgamated responses be explained (verifiability and completeness vs. accuracy levels)?

What are some of the reasons for the significant variations in quality of answers across Q&A sites?

The fact that Q&A sites are formed around different communities and use information and communication technologies in different ways (Rosenbaum & Shachaf, in press) may explain the variations in answer quality found in this study. Answer quality is significantly different across various Q&A sites probably as a result of variations across Q&A sites in community size, user demographics (e.g., age, gender, education level), policies and training mechanisms, motivators and technology infrastructure and use, as well as possible variations in question types on each of these sites. For example, policies on Yahoo! Answers are set in a top-down process but they are developed bottom-up on the Wikipedia Reference Desk (Rosenbaum & Shachaf, in press). This type of mass participation on the Wikipedia Reference Desk may be appealing for committed users who have higher interest in the success of the Q&A site compared with Yahoo! Answers users, who are expected to comply with the top down approach.

Other community features that might affect the variations in answer quality across Q&A sites include the role of the site as a component in members' identity. Wikipedia users are identified with the larger Wikipedia community and not specifically with the Wikipedia Reference Desk (Rosenbaum & Shachaf, in press); the larger Wikipedia community is engaged in mass knowledge production where information accuracy, reliability, and verifiability are major issues. The other Q&A sites foster user identification with each of the Q&A sites but not with the larger communities of their parent organizations; user profiles on these three Q&A sites indicate user' activities and ranks on the Q&A sites. Furthermore, additional demographic variations (e.g., age, gender, education level) between the Wikipedia Reference desk and the other Q&A sites may be significant. For example, while the Wikipedia Reference Desk is a male dominated environment (Shachaf, 2009), Q&A sites are female dominated as they attract mainly stay at home moms and teenagers.

In terms of community size, among the four Q&A sites, the Wikipedia Reference Desk is probably the least popular site; it does not attract as many questions as Yahoo! Answers, WikiAnswers, Answerbag, or even Askville (Hitwise, 2008). One could assume that high traffic (many questions and answers) may reflect high level of user satisfaction, which in-turn may indicate high level of answer quality and may correlate with high quality of information (Ong, Day, & Hsu, 2009). However, it is interesting that the quality of answers on Yahoo! Answers is inferior to the other less popular Q&A sites and at the same time, the Wikipedia Reference Desk provides the highest answer quality while having the lowest number of questions.

Finally, the variations across Q&A sites in their technological infrastructure and their built-in incentive mechanisms may account for the variations in performance levels. It is clear that a better understanding of these features of the various online Q&A communities could shed light on some of the reasons for the variations found in answer quality; future research into these communities is much needed.

How can the mixed findings about the quality differences between "best answers" and amalgamated responses be explained?

By comparing best answers and amalgamated responses, the argument that the crowd may produce high quality answers to questions that may resemble or surpass the quality that individuals can provide (Shachaf, 2009) was partially examined here. This argument was based on the expectation that mass collaboration and information sharing can result in the provision of cost effective high service quality for and by prosumers (Tapscott & Wiliams, 2006). One of the reasons that MIS researchers are interested in Q&A sites is that they see the potential of these systems for customer support (Ong, Day, & Hsu, 2009). However, testing this argument here (hypothesis four) generated mixed findings, which should be explained.

An amalgamated response improves the amount of verifiable information compared to a single answer, and addressed multi-parts questions better than a single answer. However, while it was possible for an amalgamated response to be more accurate than a typical single answer it was not more accurate than a single "best answer." The best answer was chosen as the best one among all the other, more typical, answers and therefore this answer possessed qualities that were better than any other answer. An amalgamated response may include, in addition to the best (and most accurate) answer, answers that are inaccurate, conflicting, or contradicting, which can cause confusion. These multiple answers that are part of an amalgamated response form a forest of mediocrity and confuse users in their quality judgments (Shachaf, 2009). Thus, an amalgamated response may be more accurate than a typical answer or a first answer on Q&A sites that do not rank best answers, such as the Wikipedia Reference Desk and WikiAnswers, but "best answers" are more accurate than an amalgamated response. Future research can examine if an amalgamated response is better than the first answer on these sites. Future research can also examine how many answers are needed to reach a critical mass for a response to be accurate and at what point additional information may potentially reduce answer quality; it should try to identify how many answers are optimal for highest quality.

CONCLUSION

While most research on Q&A sites to date has focused on Yahoo! Answers and has rarely examined more than one Q&A site at a time, this study reports on a large scale comparative analysis of answer quality on four Q&A sites. The findings indicate that answer reliability varies on Q&A sites, and provide evidence that the Wikipedia Reference Desk answers are more accurate and verifiable than answers on Yahoo! Answers, WikiAnswers, and Askville. The most popular Q&A sites, Yahoo! Answers and WikiAnswers provide answers with the lowest level of verifiable information. Amalgamated responses on Yahoo! Answers are significantly better than "best answers" in terms of completeness and verifiability and completeness but are not more accurate than the "best answers".

The findings indicate that crowd-sourcing should be implemented with great caution, because while some aspects of costumer support can be improved, improvement is not guaranteed across all service dimensions. Q&A sites are popular; yet, they provide information that may not resemble the quality of information that professionals (i.e., reference librarians) are providing. The Wikipedia Reference Desk seems to be an outlier among these Q&A sites, and the potential benefits of amalgamated answers are questionable. A Q&A site's popularity does not necessarily correlate with the quality of information it provides users.

REFERENCES

1. Adamic, L.A., Zhang, J., Bakshy, E. and Ackerman, M.S. (2008) Knowledge sharing and Yahoo! Answers: Everyone knows something, in *Proceedings of the International World Wide Web Conference*, Beijing, ACM.
2. Blooma, J.M., Chua, A.Y.K. and Goh, D.H. (2008) A predictive framework for retrieving the best answer, in *Proceedings of the 2008 ACM Symposium on Applied Computing*, Fortaleza, Ceara, Brazil, ACM.
3. Gazan, R. (2006) Specialists and synthesists in a question answering community, in *Proceedings of the American Society for Information Science & Technology Annual Meeting,* 43, 1, 1-10.
4. Giles, J. (December, 2005) Internet encyclopedias go head to head, *Nature,* December 14, 2005. Retrieved August 19, 2008 from http://www.nature.com/news/2005/051212/full/438900a.html
5. Harper, F.M., Raban, D., Rafaeli, S. and Konstan, J.A. (2008) Predictors of answer quality in online Q&A sites, in *Proceedings of the Conference on Human Factors in Computing Systems*. Florence, ACM.
6. Hitwise (2008, March 19) *U.S. visits to question and answer websites increased 118 percent year-over-year,* Retrieved November 25, 2009 from http://www.hitwise.com/news/us200803.html
7. Keen, E. (2008) The cult of the amateur: How today's Internet is killing our culture, Doubleday/Currency, UK.
8. Krippendorff, K. (2004) *Content analysis: An introduction to its methodology*, Sage, Thousand Oaks, CA.
9. Ong, C. Day M. and Hsu, M. (2009) The measurement of user satisfaction with question answering systems, *Information & Management,* 46, 7, 397-403.
10. O'Neill, N. (2007) Chacha, Yahoo!, and Amazon, *Searcher*, 15, 4, 7-11.
11. Rosenbaum, H. and Shachaf, P. (in press) A structuration approach to online communities of practice: The case of Q&A communities, *Journal of the American Society of Information Science and Technology*.
12. Shachaf, P. (2009) The paradox of expertise: Is the Wikipedia Reference Desk as good as your library? *Journal of Documentation, 65(6),* 977-963.
13. Shachaf, P. and Rosenbaum, H. (2009) Online social reference: A research agenda through a STIN framework. *Proceedings of the iConference 2009*, Feb 8-11, 2009, Chapel Hill, NC.
14. Surowiecki, J. (2004) *The wisdom of crowds*. New York: Anchor Books.
15. Weinberger, D. (2007) *Everything is miscellaneous: The power of the new digital disorder.* Henry Holt & Co, NY.