Effects of Expertise Differences in Synchronous Social Q&A

Ryen W. White
Microsoft Research
Redmond, WA 98052
ryenw@microsoft.com

Matthew Richardson
Microsoft Research
Redmond, WA 98052
mattri@microsoft.com

ABSTRACT
Synchronous social question-and-answer (Q&A) systems match askers to answerers and support real-time dialog between them to resolve questions. These systems typically find answerers based on the degree of expertise match with the asker's initial question. However, since synchronous social Q&A involves a dialog between asker and answerer, differences in expertise may also matter (e.g., extreme novices and experts may have difficulty establishing common ground). In this poster we use data from a live social Q&A system to explore the impact of expertise differences on answer quality and aspects of the dialog itself. The findings of our study suggest that synchronous social Q&A systems should consider the relative expertise of candidate answerers with respect to the asker, and offer interactive dialog support to help establish common ground between askers and answerers.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**] – Information Search and Retrieval: *search process*, *selection process*.

Keywords
Synchronous social Q&A; Expert finding; Expertise location.

1. INTRODUCTION
Synchronous social question answering (Q&A) systems match askers to answers in real time and facilitate dialog between them, usually via instant messaging (IM) [3][9]. These systems typically use expert finding methods (e.g., [1]) to identify candidate answerers based on the match between their profile and the initial question. However, since synchronous social Q&A systems initiate direct dialog between askers and answerers, the ability of askers and answerers to converse with each other effectively may also affect the outcome. Indeed, psychologists and human-computer interaction researchers have shown that expertise differences can hinder dialog between domain novices and experts [4][6] and that IM is a difficult medium for establishing the common ground that is important in successful dialog [5]. Understanding the effect of expertise differences on synchronous Q&A can help in designing better social Q&A systems, perhaps by including relative expertise when selecting candidate answerers [7], or providing support to help ground the Q&A dialog.

In this poster we present a study of expertise differences in synchronous social Q&A. We use data from a live synchronous social Q&A system, *IM-an-Expert*, deployed to a community of over two thousand users [9]. IM-an-Expert receives questions via IM, identifies candidate answerers by ranking all users according to the match between their expertise and the question, routes questions only to those available to answer, and mediates the dialog between the asker and answerer. In IM-an-Expert, all users ask or answer questions; to ask, users must be willing to answer. We use the data from the system to study the effect of expertise differences on answer ratings (assigned by askers and indicating answer quality) and the dialog.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

2. IM-AN-EXPERT
The asker poses the question to IM-an-Expert via IM. The system selects (using BM25 [8]) users who seem to be most likely able to answer a question using profile information from three sources: (i) *self-reported knowledge:* a Web interface lets users create and update a profile comprising keywords and URLs of pages about them or their interests; (ii) *email sent to mailing lists:* archives of email sent to internal mailing lists, available on a range of topics. Email is preprocessed to exclude headers and quoted text so that each profile contains only the user's authored text, and; (iii) *history*: each user is associated with questions they answered, allowing the system to improve over time. Other factors such as elapsed time since a user was last asked a question are also considered. A small group of experts who are currently available (not busy, away, etc. via presence data from the IM client) are contacted via IM three at a time, in descending order of expertise, to determine whether they can help answer the question. If and when an answerer accepts, other requests are canceled. If a candidate answerer does not respond in time or rejects the question, the service asks others. Once an answerer accepts, IM-an-Expert mediates the conversation. When the conversation ends, the asker can optionally rate answer quality from one (not helpful) to five (very helpful).

3. STUDY

3.1 Data
IM-an-Expert is deployed within Microsoft and thousands of employees use it to find answers to their questions. We use a set of 1,725 questions from 937 users. 1,144 (66%) of the questions were answered, of which 908 (79%) were rated. 527 users asked at least one question and 573 users answered at least one question. For each question we have the text, the asker and answerer identity, and the full-text of the IM dialog. We also have the profiles of askers and answerers, not including questions that they answered well ((iii) above), which varied based on *when* the question was asked and excluding it simplified our analysis. On average, users provided 8.0 unique keywords and 2.7 unique URLs. The average profile length was 8,930 words (after removing HTML tags).

3.2 Estimating Expertise
To study the effect of differences in the expertise levels on the question outcomes and the dialog itself, we devised a way to estimate the expertise of the asker and the answerer with respect to the question. Following removal of stopwords from the question and from user profiles, we represent each as term vectors with term frequency counts. We then compute the cosine similarity between the question and the profile. We do this on a per-question basis since expertise may vary by question topic. Cosine was used since it gave us a normalized measure of expertise, in the range [0,1], that could be computed for both askers and answerers and easily compared to estimate expertise differences. It also has the advantage of estimating expertise relative to the question at hand, rather than using measures such as reputation or answer history, which are question independent. We use the difference (d) between the similarities of the asker and the answer as a proxy for differences in expertise. Reassuringly, the results show that for 76% of questions, the answerer is at least as expert as the asker

and in most cases (58%), is more expert. It is worth noting that using the cosine similarity as the measure of expertise penalizes people whose profiles span multiple areas of expertise. Since such profiles may indicate less expertise in the question than someone whose profile focuses only on that topic, it is unknown whether this is a concern. We could also have focused on asynchronous media such as stackoverflow.com or quora.com. However, targeting synchronous Q&A better aligned with our research focus and let us study dialog dynamics.

3.3 Findings

We focused our analysis of the effect of expertise differences on two important aspects of the synchronous social Q&A process: (i) outcome (answer rating) and (ii) dialog (balance of conversation).

3.3.1 Expertise Difference and Answer Ratings

We were interested in the relationship between the difference in expertise and the quality of the answer. Answer quality is a measure of the benefit to the asker of the dialog. The top row of Table 1 presents the average answer rating depending on whether the answerer was less, equally, or more expert than the asker.

Table 1. Average answer rating and percentage of dialog from question asker, given expertise differences. *Ask*=asker, *Ans*=answerer.

Metric	Expertise differences		
	Ans < Ask	***Ans = Ask***	***Ans > Ask***
Average answer rating	3.66	4.05	4.27
% dialog from asker	54.8	54.7	59.3

Table 1 shows differences in the answer rating as answerers become more expert than askers. All differences in ratings are statistically significant with an independent measures analysis of variance (ANOVA), ($F(2,905) = 4.63$, $p = .01$) and post-hoc testing as appropriate. Figure 1 shows a more granular breakdown of rating by d. Error bars denote standard error of the mean (SEM).

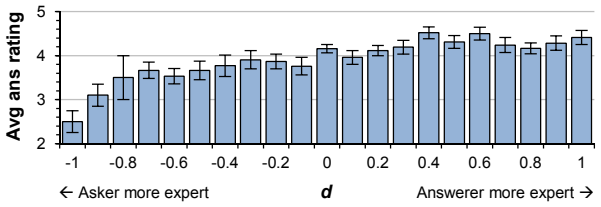

Figure 1. Average answer rating given expertise differences (± SEM).

Figure 1 shows that there is a slight increase in answer rating as answerers become increasingly more expert than askers. However, we can make more interesting observations about the figure:

1. It is not until the answerer is significantly less expert than the asker ($d \leq -0.8$) that the outcome suffers significantly. The rating is pretty robust to expertise differences (ranging from 3.7 at $d = -0.7$ to 4.4 at $d = 1.0$). One possible explanation for this is that the ability to clarify and discuss the problem via IM may lessen the importance of the expertise differences.
2. The asker just needs to have a level of expertise equivalent to the answerer to get an average answer rating of four or more.
3. More expertise is not always valuable. There is no additional asker benefit (at least in terms of answer rating) after $d \geq 0.4$. One possible explanation for this is that askers with little or no expertise on their question topic (required for such high values of *d*) may be easily satisfied with any help received, regardless of how beneficial the help was. Since our measure of answer quality is subjectively provided by the asker, we cannot distinguish between an asker receiving genuinely better answers vs. an asker who is more easily satisfied with mediocre answers.

The third observation in particular motivated us to explore the Q&A process itself for an alternative objective measure. We focus on the dialog balance, representing the relative fraction of the dialog that askers and answerers perform. This can be viewed as a measure of the fraction of the work (the relative cost) to each party of engaging in the conversation.

3.3.2 Expertise Difference and Dialog Balance

To measure the effect of expertise differences on the balance of the conversation, we computed the number of messages from each user for each IM dialog. The average percentage of each dialog coming from the asker, given differences in expertise levels, is shown in Table 1. Large differences are observed when answerers have more topic expertise than the asker ($F(2,1141) = 6.24$, $p < .01$, and $p < .02$ with Tukey post-hoc tests). One possible explanation for this phenomenon is that the asker is spending more time clarifying their question. Possible reasons for this include differences in vocabulary between novices and experts [2] and the tendency of novices to underspecify their goals in dialog with experts [6], both of which may result in more messages. In this case, it may be beneficial if the Q&A system was a more active intermediary and suggested strategies to establish common frames of reference, something that is important in successful novice-expert dialog [4]. Alternatively, it may be the case that the answerer, by being much more expert, is able to rapidly answer the question.

4. CONCLUSIONS AND FUTURE WORK

We have demonstrated the effect of expertise differences on the outcomes and dialog in synchronous social Q&A. As expected, we showed that when askers engage in dialog with those who are more expert on the question topic, they are more satisfied with the answer they receive. However, the benefit does not seem to increase beyond a certain expertise difference. In fact, beyond a difference of 0.4, the average answer rating no longer increases yet the dialog continues to become more and more imbalanced. Synchronous social Q&A systems should consider both asker and answerer topic knowledge during expert finding. In these systems it may be worth finding answerers slightly more knowledge than the asker, and reserve the most expert answerers for the most knowledgeable askers. There are opportunities for future work in better understanding the underlying causes of the observed dialog imbalance, in determining whether the imbalance is positive or negative, and in developing mitigation strategies if negative.

REFERENCES

[1] Balog, K. *et al*. (2006). Formal models for expert finding in enterprise corpora. *SIGIR*, 43-50.

[2] Falzon, P. (1991). Cooperative dialogues. In Rasmussen, J., Brehmer, B. and Leplat, J. *Distributed Decision Making: Cognitive Models for Cooperative Work*, 145-189.

[3] Horowitz, D. and Kamvar, S.D. (2010). The anatomy of a large social search engine. *WWW*, 431-440.

[4] Issacs, E.A. and Clark, H. (1987). References in conversations between experts and novices. *J. Exp. Psych.*, 116(1): 26-37.

[5] McCarthy, J.C *et al*. (1991). An experimental study of common ground in text-based communication. *SIGCHI*, 209-215.

[6] Pollack, M.E. (1985). Information sought and information provided: an empirical study of user/expert dialogues. *SIGCHI*, 155-159.

[7] Smirnova, E. and Balog, K. (2011). A user-oriented model for expert finding. *ECIR*, 580-592.

[8] Spärck-Jones, K. *et al*. (2000). A probabilistic model of information retrieval. *IP&M*, 36(6): 779-840.

[9] White, R.W. *et al*. (2011). Effects of community size and contact rate in synchronous social Q&A. *SIGCHI*, 2837-2846.