Signals of Expertise in Public and Enterprise Social Q&A

Shimei Pan
University of Maryland, Baltimore County
Baltimore, MD 21250
shimei@umbc.edu

Elijah Mayfield
LightSIDE Labs
Pittsburgh, PA 15232
elijah@lightsidelabs.com

Jie Lu and **Jennifer Lai**
IBM Research
Yorktown Heights,NY 10598
{jielu, jlai}@us.ibm.com

Abstract

The success of internet social question-answering (Q&A) sites has led businesses to attempt to duplicate this model internally. In such systems, demand for expertise is high, meaning that finding high-quality answers quickly is a top priority. Prior work on expertise discovery, though, has focused on high-traffic public websites like StackExchange. Here, we show that the most predictive metadata features from public websites do not transfer well to the enterprise.

Introduction

Information seeking online is now a social task, as online communities have shifted to a focus on user-generated content. Collaborative content creation has opened up rapid access to expertise, including information that search engines and knowledge bases alone do not provide. This includes personalized explanation of difficult concepts, instructions for completing uncommon tasks, or personal advice. Social question answering sites, like Quora, StackExchange, and Yahoo! Answers have been linchpins of that user-driven model (Anderson, Huttenlocher, and Kleinberg 2012). Many businesses wish to duplicate this model internally, enabling fast information exchange (Karimzadehgan, White, and Richardson 2009), as the ability to connect employees efficiently is crucial in business (Campbell et al. 2003). However, in the large body of research that has defined the characteristics of expert answers, almost all work has studied large and high-traffic public websites. A key assumption, then, is that the results of exploration in that domain will transfer to a corporate environment. If that domain transfer is uncertain, it limits the generality of findings for enterprise social software development.

In this paper, we explore domain transfer in expertise finding using two similar datasets: one from a high-traffic public website StackExchange[1] and one from an internal corporate social Q&A system IBM Answers. For StackExchange, we chose to exclusively sample from the Webmasters site since its topics are thematically similar to the questions asked within IBM Answers. Each source contains thousands of questions and a large base of active users. Both present questions and answers similarly (see Figure 2), and their statistics are roughly comparable, though the public site has higher traffic (see Table).

Feature Clusters from Prior Work

To systematize findings from prior work, we reviewed approaches that have been used for characterizing answer quality in social Q&A in public datasets. Our goal was to identify high-level views that had been proven effective in public datasets; these are likely to motivate enterprise developers when following due diligence in building their own systems. Overall, we have identified five loosely-grouped approaches to measuring answer quality. Each corresponds to a high-level hypothesis about the "nature" of what produces high-quality answers in a social Q&A context. These hypotheses are presented in Figure 1. The features chosen to represent those clusters are presented in Table 2.

Asker Cluster Features in this cluster represent metadata about the author of the question being replied to. They correspond conceptually to **H1** in Figure 1, which states that the reason a question receives high-quality answers is because there is some characteristic of the question asker that leads to quality.

Answerer Cluster This cluster represents metadata about the author of the answer being judged for quality. They relate to **H2**; this hypothesis claims simply that certain users write answers that are consistently judged as high quality. These features are entirely based on activity and on-site behavior, rather than observing content or domain expertise in answers.

Context Cluster This cluster represents a counterfactual - that traffic, rather than any characteristics of an answer, are what drives our measures of answer quality. Because number of votes in particular is likely to increase with more users interacting with a given thread of answers to a question, this may be artificially inflating our values. This corresponds to our hypothesis **H3**.

Shallow Text Cluster This cluster represents the most pessimistic view of answer quality - namely, that simple heuristics about an answer's text, such as the number of words or the amount of formatting, can predict quality of

[1] www.stackexchange.com

H1 (Asker). Particular users are more likely to ask questions that receive high-quality responses.

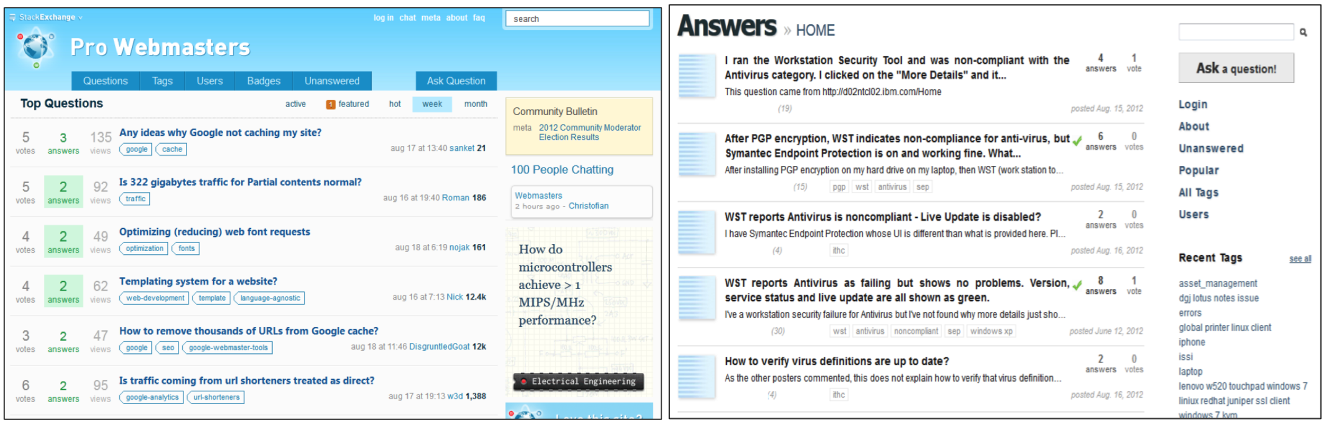
H2 (Answerer). Particular users are more likely to give answers that are judged as high-quality.

H3 (Context). Measures of answer quality are affected primarily by that answer's context and a question's traffic.

H4 (Shallow Text). Simple metrics of textual content, such as length, can distinguish high-quality answers.

H5 (Shallow Style). Stylistic features of text, such as pronoun usage, can distinguish high-quality answers.

Figure 1: Statement of high-level hypotheses driving our grouping of features into clusters, based on findings from prior work.



5 votes 3 answers 135 views **Any ideas why Google not caching my site?** google cache aug 17 at 13:40 sanket **21**

5 votes 2 answers 92 views **Is 322 gigabytes traffic for Partial contents normal?** traffic aug 16 at 19:40 Roman **186**

4 votes 2 answers 49 views **Optimizing (reducing) web font requests** optimization fonts aug 18 at 6:19 nojak **161**

4 votes 2 answers 62 views **Templating system for a website?** web-development template language-agnostic aug 16 at 7:13 Nick **12.4k**

3 votes 2 answers 47 views **How to remove thousands of URLs from Google cache?** google seo google-webmaster-tools aug 18 at 11:46 DisgruntledGoat **12k**

6 votes 2 answers 95 views **Is traffic coming from url shorteners treated as direct?** google-analytics url-shorteners aug 17 at 19:13 w3d **1,388**

Community Bulletin
meta 2012 Community Moderator Election Results

100 People Chatting

Webmasters
2 hours ago - Christofian

How do microcontrollers achieve > 1 MIPS/MHz performance?

I ran the Workstation Security Tool and was non-compliant with the Antivirus category. I clicked on the "More Details" and it... 4 answers 1 vote
This question came from http://d02ntcl02.ibm.com/Home
(19) posted Aug. 15, 2012

After PGP encryption, WST indicates non-compliance for anti-virus, but Symantec Endpoint Protection is on and working fine. What... ✓ 6 answers 0 votes
After installing PGP encryption on my hard drive on my laptop, then WST (work station to...
(15) pgp wst antivirus sep posted Aug. 15, 2012

WST reports Antivirus is noncompliant - Live Update is disabled? 2 answers 0 votes
I have Symantec Endpoint Protection whose UI is different than what is provided here. Pl...
(4) ithc posted Aug. 16, 2012

WST reports Antivirus as failing but shows no problems. Version, service status and live update are all shown as green. ✓ 8 answers 1 vote
I've a workstation security failure for Antivirus but I've not found why more details just sho...
(30) wst antivirus noncompliant sep windows xp posted June 12, 2012

How to verify virus definitions are up to date? 2 answers 0 votes
As the other posters commented, this does not explain how to verify that virus definition...
(4) ithc posted Aug. 16, 2012

asset_management
dgj lotus notes issue
errors
global printer linux client
iphone
issi
laptop
lenovo w520 touchpad windows 7
liniux redhat juniper ssl client
windows 7 kvm

Figure 2: Screenshots from the public StackExchange website (left) and the internal corporate website (right) used in our study.

Statistic	Public	Enterprise
Total # Questions	8674	3968
Total # Answers	16020	4134
Total # Users with 1+ Posts	6004	3352
Mean # Answers / Question	1.85	1.04
Mean # Questions Posted Per User	1.44	1.18
Mean # Answers Posted Per User	2.67	1.23
Mean # Votes Per Answer	1.95	0.54
Answers Accepted (%)	24.6	25.6
Questions w/ Accepted Answer (%)	45.5	26.6

Table 1: Dataset statistics for both datasets

that answer, without deeply analyzing that text. This is our hypothesis **H4**.

Shallow Style Cluster This cluster represents the linguistically motivated features tied to the belief that answer quality judgment is somewhat separated from the content of the answer or the personalities of the users. We draw particularly on three bodies of work - one claiming that signals of trust and hedging language predicts answer quality (Su, Huang, and Chen 2010), one showing that first-person pronouns predict answer quality (using the Linguistic Inquiry and Word Count dictionary (Tausczik and Pennebaker 2011)), and one showing the same for positive and negative emotional affect (Aji and Agichtein 2010). Though their hypotheses are disparate, we group them into hypothesis **H5**.

Methods

Measuring Answer Quality

Our analysis of answer quality uses two possible methods of evaluation, differing in whose satisfaction is being measured. Both judgments are clearly visible in the user interfaces of StackExchange and IBM Answers.

1. **Accepted Answer** - we assign a score of 1 to answers that were accepted by the information seeker, and a score of 0 to answers that were not accepted (meaning, at most, only one answer to each question can receive any credit). This is a direct measure of the question *asker*'s satisfaction with an answer.

2. **Community Votes** - An answer's score is the number of net votes an answer receives (positive votes minus negative votes). This metric gives credit to any answer that receives votes, even if it is not accepted by the question asker. It is a proxy for the evaluation of the larger set of users visiting a website, rather than the original question's author, and therefore represents the satisfaction of question *readers*.

Experimental Setup

From our data, 2,000 answers were randomly sampled, using 1,000 answers from each site. When sampling we did not limit our answers to distinct questions; the answers in this subset therefore correspond to 1,922 distinct questions. Then, for each feature cluster we described, we extracted several features. All have been proven effective in prior

Asker Cluster
Number of questions this user has asked
Total activity of a user (number of questions, answers, and comments)
Percent of previous questions from this user with an accepted answer
Time elapsed since user's first post
Answerer Cluster
Total activity of a user (number of questions, answers, and comments)
Total votes received from past activity
Percentage of previously written answers which were accepted by the question asker
Average votes received per day since joining
Total community votes received in the last 7 days
Context Cluster
Number of answers the question received
Total votes for all posts (questions, answers, and comments) related to a question.
Number of comments responding to the highest-voted answer to the question
Time elapsed between question being posted and this answer being posted.
Answers already posted before this answer
Shallow Text Cluster
Number of words in a response
Number of overlapping words (excluding stopwords) between question and answer
Number of unique terms only present in answer, not question
Number of HTML formatting tokens in an answer
Number of HTML hyperlinks in an answer
Shallow Style Cluster
First-person singular pronoun ratios *(I, me, my)*
First-person plural pronoun rations *(we, us, our)*
"Cognitive Process" LIWC keywords *(cause, ought, know)*
Number of positive emotion keywords
Number of negative emotion keywords
"Absolute-trust" keywords *(definite, sure)*
"High-trust" keywords *(clearly, obviously)*
"Moderate-trust" keywords *(seemingly, should)*
"Low-trust" keywords *(doubt, perhaps, might)*

Table 2: Features in each cluster from prior work.

work. We extracted the same features from both private and public datasets.

Then, for each dataset, we performed two multiple regressions, varying the dependent variable (one of our two measures of answer quality). We model answer acceptance as a binomial, as it has only two possible values. For each cluster, all described features were used as independent variables; we did not attempt to measure any higher-level interactions between variables.

Results

In the public domain dataset, all five clusters produced multiple regressions which were significantly[2] predictive of both answer acceptance percentage and number of votes received, with the exception of the Shallow Style cluster, which only predicted votes received. These results are summarized in Table 3. At a broad level, this verifies the findings of prior work and confirms that those indicators are indeed useful for the domain in which they were originally situated.

As we conjectured, effects differ when transferring to the corporate domain. The relative change is visualized in Figure 3. In four of five clusters, we see decreases in correlation coefficient[3]. In the Answerer cluster, on the other hand, we see a large increase in predictive power.

Several observations become apparent from deeper analysis of the coefficients assigned in these multiple regressions.

1. **Influences from Traffic**

Within the public StackExchange data, the Context cluster of features was most highly correlated with both measures of answer quality ($r = .357$ and .386 for answer acceptance and community votes, respectively, as shown in Table 3). This breaks down across quality measures, though, into opposite values in per-feature coefficients. When predicting the number of community votes received, coefficients pointed towards more existing traffic (measured in votes or comments) resulting in more votes for new answers. In this case, at least, a rising tide lifts all boats. For answer acceptance, we see the opposite effect. Answers were more likely to be accepted when there were fewer comments and fewer other questions being posted.

These features were much weaker fits for the corporate data. While the trends still existed in the data, following the same patterns, effects were muted. This is likely due to the more targeted, lower-traffic nature of the business site. We also propose, tentatively, that users are acting less strategically for online credibility measures like votes. This suggests that reputation is a less-impactful motivator for employees to contribute to an enterprise social Q&A website.

Traffic indicators are supported by game theoretic predictions. Answers are more likely to be accepted due to lack of competition; this parallels prior work (Jain, Chen, and Parkes 2009). Experts may also be more likely to respond to questions which have not already received attention from others. Users are more likely to feel like they are helping others, and are more likely to receive the reward incentive of an accepted answer, if they are not competing with a larger crowd. Both of these reasons have been suggested by prior work (Dearman and Truong 2010), and our data supports those hypotheses.

2. **Visual Formatting and Presentation**

The Shallow Text cluster of features was a highly significant predictor in the StackExchange data ($r = .295$ in prediction of community votes, as shown in Table 3). However, answer length was not the most predictive feature within that cluster. Instead, the feature measuring HTML formatting was given the most weight. Some of this can be attributed to bold and italic text, but the most common theme among highly-voted answer formatting was lists, either of bullets or numbered items, and frequent use of monospaced

[2] In this and all instances, throughout this paper, significance is defined as $p < 0.01$.

[3] It is inappropriate to test statistical significance for r values across different datasets, thus, we exclude p values for this decrease.

	Public		Enterprise	
Cluster	Accept	Votes	Accept	Votes
Asker	**.249**	**.225**	**.208**	**.212**
Answerer	**.226**	**.242**	**.265**	**.556**
Context	**.357**	**.386**	**.275**	**.149**
Shallow Text	**.170**	**.295**	.087	.115
Shallow Style	.124	**.247**	.100	.131

Table 3: Multiple regression r for each feature cluster. **Bold** correlations are significant.

fonts (usually used for presenting program code). This corresponds to findings from the MathOverflow website, which found equation formatting to be a predictor of answer quality (Tausczik and Pennebaker 2011).

These patterns were again weaker in the corporate data. This effect is conflated with domain - while the Webmasters forum from StackExchange was chosen to parallel our corporate dataset, it still had more questions requiring programming code. This means that monospaced formatting was not as prevalent in the corporate data. The correlation was much stronger for community votes than for answer acceptance, which suggests that readers pay more attention to formatting while askers pay more attention to information content.

3. Answerer Identity and Individual Behavior Patterns

The majority of features from prior work are a poorer fit for enterprise data. The Answerer cluster is an exception, explaining a high amount of variance in community votes within the enterprise ($r = .556$, as shown in Table 3), far above any other category tested. Most of this variance comes from the feature measuring the number of recent votes a user has received (based on their activity in the last 7 days). This one feature accounted for more variance in this cluster than all other features combined. By contrast, in the public dataset, a user's long-term average votes per day (since registration) is a higher predictor of answer quality. These results indicate a pattern of "burst" behavior in an enterprise.

We conjecture that in an enterprise, the burst behavior of an expert is trigged by fortuitous events that are unplanned (e.g., stumble upon a search result with links to Q&A sites). Since social Q&A is outside the routine workflow of an enterprise knowledge worker, without actively maintaining the link between an expert and a Q&A community, there is a high probability that the expert's participation drops or even completely stops after a brief period of time.

Conclusions

Our results above have demonstrated that the majority of features from previous work on public social Q&A are a poorer fit for enterprise data. To be able to detect the "burst" behavior quickly is the key to finding expertise in the enterprise. Relying on metadata however means experts cannot be discovered from infrequent but high-quality participation. These findings have implications for both enterprise social software management and researchers in organizational behavior.

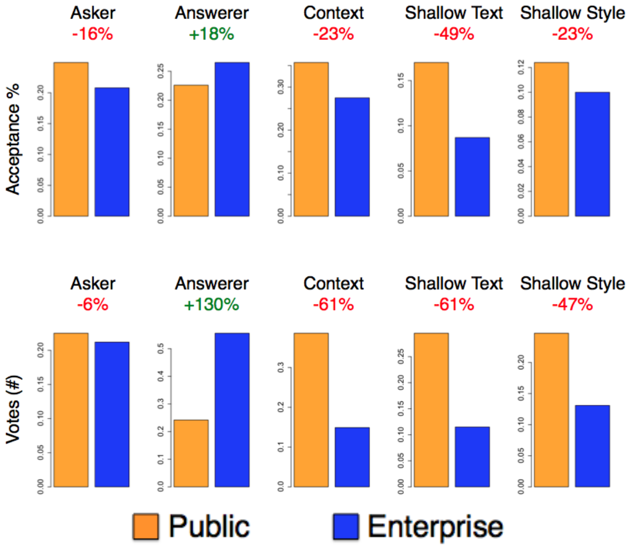


Figure 3: Changes in both measures of answer quality, per cluster, when moving to enterprise datasets. Y axes are scaled to a constant maximum height to emphasize relative gain.

References

Aji, A., and Agichtein, E. 2010. The "nays" have it: Exploring effects of sentiment in collaborative knowledge sharing. In *NAACL*.

Anderson, A.; Huttenlocher, D.; and Kleinberg, J. 2012. Discovering value from community activity on focused question answering sites: A case study of stack overflow. In *KDD*.

Campbell, C.; Maglio, P.; Cozzi, A.; and Dom, B. 2003. Expertise identification using email communications. In *CIKM*.

Dearman, D., and Truong, K. 2010. Why users of yahoo! answers do not answer questions. In *CHI*.

Jain, S.; Chen, Y.; and Parkes, D. 2009. Designing incentives for online question and answer forums. In *Proceedings of ACM Conference on Electronic Commerce*.

Karimzadehgan, M.; White, R.; and Richardson, M. 2009. Enhancing expert finding using organizational hierarchies. In *ECIR*.

Su, Q.; Huang, C.; and Chen, H. 2010. Evidentiality for text trustworthiness detection. In *Proceedings of ACL Workshop on NLP And Linguistics*.

Tausczik, Y., and Pennebaker, J. 2011. Predicting the perceived quality of online mathematics contributions from users' reputations. In *CHI*.