

Evaluating and Predicting Answer Quality in Community QA

Chirag Shah

School of Communication & Information (SC&I)
Rutgers, The State University of New Jersey
4 Huntington St, New Brunswick, NJ 08901
chirags@rutgers.edu

Jefferey Pomerantz

School of Information & Library Science (SILS)
University of North Carolina at Chapel Hill
100 Manning Dr, Chapel Hill, NC 27599
pomerantz@unc.edu

ABSTRACT

Question answering (QA) helps one go beyond traditional keywords-based querying and retrieve information in more precise form than given by a document or a list of documents. Several community-based QA (CQA) services have emerged allowing information seekers pose their information need as questions and receive answers from their fellow users. A question may receive multiple answers from multiple users and the asker or the community can choose the best answer. While the asker can thus indicate if he was satisfied with the information he received, there is no clear way of evaluating the quality of that information. We present a study to evaluate and predict the quality of an answer in a CQA setting. We chose Yahoo! Answers as such CQA service and selected a small set of questions, each with at least five answers. We asked Amazon Mechanical Turk workers to rate the quality of each answer for a given question based on 13 different criteria. Each answer was rated by five different workers. We then matched their assessments with the actual asker's rating of a given answer. We show that the quality criteria we used faithfully match with asker's perception of a quality answer. We furthered our investigation by extracting various features from questions, answers, and the users who posted them, and training a number of classifiers to select the best answer using those features. We demonstrate a high predictability of our trained models along with the relative merits of each of the features for such prediction. These models support our argument that in case of CQA, contextual information such as a user's profile, can be critical in evaluating and predicting content quality.

Categories and Subject Descriptors

H.3: INFORMATION STORAGE AND RETRIEVAL H.3.3: Information Search and Retrieval: *Search process*; H.3.5: Online Information Services: *Web-based services*

General Terms

Experimentation; Human Factors; Measurement.

Keywords

Community question answering; answer quality evaluation and prediction.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR '10, July 19–23, 2010, Geneva, Switzerland. Copyright 2010 ACM 978-1-60558-896-4/10/07...\$10.00.

1. INTRODUCTION

Community Question Answering (CQA) sites have emerged in the past few years as an enormous market, so to speak, for the fulfillment of information needs. Estimates of the volume of questions answered are difficult to come by, but it is likely that the number of questions answered on CQA sites far exceeds the number of questions answered by library reference services [4], which until recently were one of the few institutional sources for such question answering. CQA sites make their content – questions and associated answers submitted on the site – available on the open web, and indexable by search engines, thus enabling web users to find answers provided for previously asked questions in response to new queries.

Yahoo! Answers¹ and AnswerBag² are examples of such CQA services, the popularity of which have been increasing dramatically for the past several years.³ The fact that CQA sites receive such a high volume of use, and that there is such a large ocean of information needs that may be fulfilled by these sites, makes it critical to establish criteria for evaluating the quality of answers provided by these sites. Library reference services have a long tradition of evaluation to establish the degree to which a service is meeting user needs [9]. Such evaluation is no less critical for CQA sites, and perhaps even more so, as these sites do not have a set of professional guidelines and ethics behind them, as library services do.

A small but growing body of literature exists on evaluating the quality of answers provided on CQA sites. Some work has been done to predict asker satisfaction [8], but little work exists on what factors contribute to a quality answer, beyond simply being satisfactory to the asker.

As Liu et al. [8] point out, research on CQA draws on research on interactive Question Answering. One significant distinction, however, is that in studying CQA there is an actual user, while the questions used in the TREC QA track, for example, are questions submitted to the FAQ Finder system, and the found answers (by participating sites) are evaluated by trained assessors [15]. It is of course appropriate to reduce the subjectivity of evaluating answers in a large-scale evaluation of systems like a TREC track.

¹ <http://answers.yahoo.com>

² <http://www.answerbag.com>

³ According to a March 2008 Hitwise report (<http://www.hitwise.com/press-center/hitwiseHS2004/question-and-answer-websites.php>), visits to CQA websites have increased 118% year-over-year since 2006.

However, in CQA, the subjectivity of relevance assessments is central. Furthermore, in evaluations of CQA, it becomes possible to glean relevance assessments from actual users – rather than trained assessors – possibly even the asker’s own relevance assessments.

Beyond the benefit to the user of having better metrics for evaluating the quality of answers, it would benefit the management of CQA sites themselves to have such metrics. Many CQA sites have implemented reputation systems, where answerers can gain points, or advance levels, based on their participation on the site: number of questions answered, number of answers voted as best answers, etc. Having metrics for evaluating the quality of an answer would enable CQA sites to add this as a factor in building an answerer’s reputation. A reputation system that incorporated such quality metrics would also benefit askers: an asker could view the profiles of answerers who provided answers to his question, and see the quality of past answers provided by that answerer.

Some researchers suggest that the context that gives rise to an information need is unique for every individual, and that an answer that is useful to an individual in a particular context will therefore be at best only partially useful to others in other contexts [1]. Identifying the factors that contribute to the quality of answers would therefore be critical for informing web users in deciding if a previously provided answer is appropriate for answering their question.

In this paper we present a novel approach to measuring the quality of answers on CQA sites, and use it to predict which of the answers to a given question the asker will pick as the best answer. The specific problem of answer quality prediction is defined in Section 3. This is followed by our two approaches to measure the quality of the answers. In Section 4, we discuss how we used human assessment of different aspects of answer quality, and in Section 5, we show how we used automatically extracted features for measuring and predicting quality of answers. An overview of some of the related work is provided in the following section.

2. BACKGROUND

Community Question Answering, according to Shah et al. [13], consists of three components: a mechanism for users to submit questions in natural language, a venue for users to submit answers to questions, and a community built around this exchange. Viewed in that light, online communities have performed a question answering function perhaps since the advent of Usenet and Bulletin Board Systems, so in one sense CQA is nothing new. Websites dedicated to CQA, however, have emerged on the web only within the past few years: the first CQA site was the Korean Naver Knowledge iN, launched in 2002, while the first English-language CQA site was Yahoo! Answers, launched in 2005. Despite this short history, however, CQA has already attracted a great deal of attention from researchers investigating information seeking behaviors [5], selection of resources [3], social annotations [2], user motivations [12], comparisons with other types of question answering services [14], and a range of other information-related behaviors.

While CQA sites are new, they have been in existence long enough for product differentiation to begin to occur. Many CQA sites allow questions to be submitted on any topic: for example, Yahoo! Answers, WikiAnswers, AnswerBag, and others. Several CQA sites, however, restrict their scope in a variety of ways.

Some sites are subject-specific, such as Stack Overflow, which limits its scope to questions about programming,⁴ and Math Overflow, which limits its scope to research level math questions.⁵ Some sites serve a specific user community, such as HeadHunterIQ, which targets business recruiters.⁶ Some sites answer only specific types of questions, such as Homework Hub, which is limited to homework help.⁷ From the standpoint of user satisfaction – with both the answer and the site – it would be a benefit to CQA sites for there to be a mechanism to triage questions between sites. The topic of a question would obviously be a factor in such a mechanism, but other factors in evaluating the quality of answers provided on the site could also be valuable for this purpose.

This sort of triage is comparatively simple for a human to perform – and while time-consuming, is in fact commonly performed by librarians for digital reference services [9,10]. The QuestionPoint reference service,⁸ an application for managing a global collaborative of library reference services, also performs this sort of triage automatically, by matching questions to profiles of individual libraries [6]. The level of complexity that such triage systems are currently capable of, however, pale in comparison to the complexity that a human triage can manage.

Both triage and the evaluation of the quality of answers are, of course, largely classification problems: in triage, the question must be placed in one and only one “bin” corresponding to a specific answering service, while in evaluation it must be determined if the question does or does not meet an evaluation criterion. Most classification tasks are sufficiently complex that, by and large, humans are able to perform them more effectively – if not efficiently – than machine classification. To improve machine classification, of course, a sufficiently large dataset and a sufficiently rich set of factors are needed. The work reported here investigates two methods for developing sets of factors: first, a set of 13 quality metrics identified from the literature on CQA are used for human quality assessments, then a set of features are automatically extracted.

CQA presents interesting problems for automatically extracting criteria to guide classification. Automatic classification must of course make use of a corpus of some kind. For CQA, this corpus consists of the questions and their associated answers submitted on the site, and could also include other data such as user profiles, answer ratings, etc. There is, however, much more going on CQA sites than just answering questions and reputation building. CQA sites are communities, however loosely defined, and so there is a wealth of contextual information about that community, and the interpersonal dynamics within it, that could be utilized, if only we understood how to make use of it.

In an article mapping out a research agenda for CQA, Shah et al. [13] identify two primary threads in this work: research on the content of CQA sites, which includes analyses of the content and quality of the questions and answers, and research on the community around these sites, which includes analyses of the reputation systems and social norms on these sites. While, of

⁴ <http://www.stackoverflow.com>

⁵ <http://www.mathoverflow.net>

⁶ <http://www.headhunteriq.com>

⁷ <http://www.stackexchangesites.com/homework-hub>

⁸ <http://questionpoint.org>

course, the content of a CQA site would not exist without the community, here we focus only on the former.

CQA sites may be mined to identify the information needs within a specific domain [7], but Yahoo! Answers (YA) enables questions to be asked on nearly any possible topic. This broad diversity of topics, as well as the diversity of users asking questions, makes evaluating answers challenging. Indeed, Shah et al. [12] found that while 95% of questions that they investigated on YA received at least one answer, the quality of those answers varied widely, from correct answers to partial and unhelpful answers, to spam. Further, as in any instance of information retrieval or question answering, the quality of the answer is heavily dependent user-specific factors.

One of the difficulties of studying CQA, however, is the difficulty – sometimes the impossibility – of gaining access to the members of the community themselves, askers or answerers. As a result, a range of approaches has emerged for making use of proxies for the askers themselves. Unable to contact the askers directly, Kim et al. [5] analyzed the comments provided by askers when selecting a best answer. Others have used third parties to stand in for askers: Harper et al. [3] used undergraduate students as proxies for askers, while Liu et al. [8] used both subject experts and paid workers from Amazon Mechanical Turk.⁹ Both of these approaches have clear advantages: the former makes use of the asker's own words and evaluation criteria, while the latter is able to collect more detailed evaluative data. The study presented here makes use of the latter approach.

A range of approaches has also emerged for developing evaluation criteria used in studies of CQA. Liu et al. [8] had proxies evaluate answers according to the same 5-star rating system available to the asker. Kim et al. [5] let evaluation criteria emerge from the comments provided by askers when selecting a best answer. Harper et al. [3] had proxies evaluate answers according to criteria derived from the evaluation of library reference services. Perhaps the most fully developed set of evaluation criteria for answers in CQA, however, was proposed by Zhu et al. [16], who identified a set of 13 criteria from both CQA sites' guidelines for answerers, and the comments provided by the community on questions and answers. For our work reported here, we will use these 13 criteria when asking human assessors to judge the quality of an answer.

3. PROBLEM OF ANSWER QUALITY PREDICTION

The quality of an answer, or of any information content for that matter, can be subjective. A quality assessment may depend on the relevance of that content, among other factors, and relevance itself is difficult to measure in the context of CQA. We, therefore, provide our own interpretation of quality with respect to the data and the task we have on our hand.

On YA, a question is considered to be resolved if either the community votes and selects one of the answers to be the best, or the asker himself chooses one as the best answer from the set of answers he received for his question. It is possible that multiple answers are of high quality, but only one of them gets picked as the best answer. Liu et al. [8] considered the act of an asker

choosing one as the best answer an indication of satisfaction. If the asker does not select any answer as the best one, and/or if the community votes for the best answer, the asker is assumed to be unsatisfied. For the work reported here, we will follow this notion of asker satisfaction. To extend it to indicate the quality of an answer, we will add another constraint that the asker has to give the chosen answer a rating of at least 3 out of 5. Thus, we consider an answer to be a high quality answer, if (1) the asker chose it as the best answer, and (2) gave it a rating of at least 3.

Given this, we define the problem of answer quality prediction to be one where we need to predict if a given answer will be selected by the asker as a high quality answer. Thus, the goal of our work is to predict if an answer was chosen by the asker as the best answer with a high rating. In order to do that, we will evaluate each answer's quality according to several measures. First, we will use the 13 criteria used by Zhu et al. [16] as different aspects of answer quality. Next, we will extract a number of features from questions, answers, and the profiles of their posters as a way to measure answer quality. For both of these approaches, we will demonstrate how we extracted necessary features and constructed fairly reliable models, and used these models to classify an answer to be in the 'yes' class (chosen as the best), or in the 'no' class (not chosen as the best).

4. PREDICTION MODEL USING HUMAN ASSESSMENTS

In this section we will describe how we used the quality assessment data obtained from Amazon Mechanical Turk (MTurk) human assessors (also known as Turk workers) to build a regression model for predicting answer quality. It is important to note that in a strict sense, we are not actually trying to evaluate the quality of answers. Rather, we are trying to predict if a given answer will be selected by the asker as the best answer or not. It is possible that an answer is of good quality, but was not chosen by the asker as the best answer. Conversely, an asker could select an answer as the best that another (a librarian or a subject expert, for example) would not have evaluated highly. But for the purpose of our work here, we will assume that an answer declared as the best answer by the asker with a rating of at least 3 is a high quality answer.

4.1 Data

Yahoo! Research makes several datasets available to university-affiliated researchers, through their Webscope™ Program.¹⁰ One of these datasets is Yahoo! Quest, which consists of the full text of approximately 8 million questions from Yahoo! Answers, the full text of all answers to each question, which was voted the best answer, the subject of the question as chosen by the asker, and other associated data.

In the Yahoo! Quest dataset, questions are categorized into four broad types: advice, informational, opinion, and polling. From this dataset, we randomly sampled 30 questions of each of these four types, for which there were at least five answers. For each question, we picked five answers – the answer voted as the best answer, and four others randomly sampled. Thus, our data consisted of 120 questions with 600 answers. Beyond being stratified into the aforementioned four types, the questions in our

⁹ <http://www.mturk.com>

¹⁰ <http://sandbox.yahoo.com>

dataset were quite wide-ranging, covering a broad range of subject areas and depth of subject knowledge.

4.2 Obtaining quality assessment using Mechanical Turk

As mentioned above, it is often impossible to gather evaluative data about answers from the askers themselves, and that was the case here. Users of YA may create a profile, which may include an email or IM address, but the Yahoo! Quest dataset does not include any user profile data. It was, therefore, impossible to ask the asker how they had evaluated the various answers to their question. As a proxy for the original asker, we therefore used Amazon Mechanical Turk. MTurk has been used successfully to predict asker satisfaction, and indeed ratings from Turk workers have been shown to be more closely correlated with actual asker satisfaction than ratings from experts [8].

We created a Human Intelligence Task (HIT) on MTurk, in which Turk workers evaluated each of the 600 answers in our dataset according to the 13 quality criteria identified by [16] (discussed above), on a 5-point Likert scale. In particular, we presented a worker with a question-answer pair, and instructed the worker to rate the following 13 statements on scale of 1 to 5.

1. This answer provides **enough information** for the question. (*informative*)
2. This answer is **polite** (not offending). (*polite*)
3. This answer **completely answers** the whole question (rather than a part of it). (*complete*)
4. This is an **easy to read** answer. (*readable*)
5. This answer is **relevant** to the question. (*relevant*)
6. The answer is **concise** (not too wordy). (*brief*)
7. I **trust/believe** this answer. (*convincing*)
8. The answer contains **enough detail**. (*detailed*)
9. I believe this answer is **original** (not copied from another place). (*original*)
10. I believe the answer is **objective and impartial**. (*objective*)
11. The answer has **new ideas or concepts** that made me somewhat surprised. (*novel*)
12. This answer is **useful or helpful** to address the question. (*helpful*)
13. I believe this answer has been written by an **expert**. (*expert*)

Since these are subjective criteria, it is likely that different people have different levels of agreement while rating them. We, therefore, employed five different Turk workers to obtain ratings on each of these statements. Thus, we created a total of 3000 HITs. To reduce the order effect, we randomized the ordering of the above statements for each HIT.

In order to ensure that the worker who completed our HIT was an actual human and not a bot, we added a final question: “What is two plus two,” and provided a text field for the answer. Human workers, of course, had no trouble answering this correctly, while bots supplied nonsensical responses that were easily filtered out. Workers took about 90 seconds on average to work on a HIT (read a question-answer pair and rate 13 different criteria).

4.3 Constructing a model with logistic regression

As described before, each worker on MTurk rated each of the thirteen quality aspects for an answer on scale 1 to 5. To convert this rating to a two-class decision (‘yes’ or ‘no’ for an aspect), we

assumed that a rating of 3 or higher meant ‘yes’, and a rating of 1 or 2 meant ‘no’. Thus, for a statement “*This answer contains enough details.*”, if a worker rated 3 or higher, we considered his decision to be ‘yes’ for the answer containing enough details to address the question. Since we had five different workers rating each answer, we used a simple voting scheme to arrive to their collective decision. We counted the votes for a factor to be ‘yes’ and normalized it by the number of raters (5). Thus, if a factor received 4 ‘yes’ decisions, it had 0.8 as its weight. Using such weighted aspects for each answer, we constructed a model for our dataset using logistic regression. In this model, the independent variables were the 13 aspects, and the dependent variable was the asker’s decision about the quality of an answer. As described earlier, the dataset consisted of 120 questions with total of 600 answers. Once again, an answer was considered to be a high quality answer, if (1) the asker chose it as the best answer, and (2) gave it a rating of 3 or higher.

The model constructed with logistic regression is shown in Table 1. As we can see, most factors do not contribute significantly in explaining the variability of the data. In fact, overall, pseudo R^2 is reported to be 0.0902 with $(p > \chi^2) = 0.0000$, which means only about 9% of the variability in the data could be explained by this model.

Table 1: Logistic regression model for 13 quality factors in predicting asker-rated quality of an answer.

Logistic regression		Number of obs = 600		
Log pseudolikelihood = -265.40593		wa1d ch12(13) = 55.48	Prob > ch12 = 0.0000	
		Pseudo R2 = 0.0902		
(Std. Err. adjusted for 120 clusters in qid)				
rating	Coef.	Robust Std. Err.	z P> z [95% Conf. Interval]	
informative	.4429017	.7485814	0.59 0.554	-1.024291 1.910094
polite	-.5958797	.7714024	-0.77 0.440	-2.107801 .9160413
complete	.8192924	.6996266	1.17 0.242	-.5319505 2.190535
readable	-2.118669	.8350697	-2.54 0.011	-3.753775 -.4819622
relevant	-.6858092	.7688801	-0.89 0.372	-2.192787 .8211681
brief	-.7298307	.7512585	-0.97 0.331	-2.20227 .742609
convincing	-.0764392	.8450847	-0.09 0.928	-1.732775 1.579896
detailed	1.299867	.7694943	1.69 0.091	-.2083146 2.808048
original	1.648606	.7450068	2.21 0.027	-.1884198 3.108793
objective	-1.247356	.7662476	-1.63 0.104	-2.749174 2.544616
novel	2.298834	.6831346	3.37 0.001	.9599133 3.637754
helpful	1.023509	.7079909	1.45 0.148	-.363244 2.410361
expert	-.9137472	.7107884	-1.29 0.199	-2.306867 4.793725
_cons	-2.08764	.8108309	-2.57 0.010	-3.67684 -.4984407

This indicates the low quality of the model. We do see that aspects ‘novel’, ‘original’, and ‘readable’ have significant impact on model’s ability to predict the best answer. Doing further data mining revealed that losing aspects with high $p > |z|$ values do not affect the performance of this model by much. Removing aspects ‘convincing’ and ‘informative’ reduced the power of this model by only a little, with pseudo $R^2 = 0.0896$. However, this model is not very appropriate for doing feature selection since all of the aspects are strongly correlated with each other as shown in Table 2. This may also be the reason for the model not being able to explain much of the variability in the data.

Applying this model on the same data for testing, we achieved 80.33% classification accuracy. Doing a 10-fold cross-validation with this model resulted in classification accuracy of 79.50%, with almost all of the best answers classified in the wrong class. In fact, given the kind of data we have, where 4 out of 5 answers are not selected as the best answers, we could achieve 80.00% classification accuracy by simply declaring every answer to be not the best answer. It should be noted that in reality, a question on YA may receive many more answers than this, and thus, constructing a classifier to identify the best answer is extremely difficult.

Table 2: Pair-wise correlation of aspects. Significant relations ($p < 0.05$) are indicated with ‘*’.

	inform-e	polite	complete	readable	relevant	brief	convin-g
informative	1.0000						
polite	0.3916*	1.0000					
complete	0.5369*	0.4357*	1.0000				
readable	0.2464*	0.3392*	0.2453*	1.0000			
relevant	0.4421*	0.4381*	0.4151*	0.3556*	1.0000		
brief	0.1422*	0.2348*	0.1831*	0.3634*	0.2531*	1.0000	
convincing	0.4743*	0.4612*	0.4606*	0.3789*	0.4968*	0.2868*	1.0000
detailed	0.6342*	0.4206*	0.6125*	0.2545*	0.4414*	0.1271*	0.4969*
original	0.2795*	0.3165*	0.2403*	0.3035*	0.3784*	0.3003*	0.3389*
objective	0.3651*	0.4261*	0.3996*	0.2769*	0.4560*	0.2390*	0.3812*
novel	0.4219*	0.3104*	0.4035*	0.1804*	0.3386*	0.1570*	0.3450*
helpful	0.5501*	0.4858*	0.5737*	0.3647*	0.5121*	0.2239*	0.5068*
expert	0.5144*	0.3761*	0.4650*	0.1882*	0.3491*	0.1427*	0.4198*
	detailed	original	object-e	novel	helpful	expert	
detailed	1.0000						
original	0.2413*	1.0000					
objective	0.4233*	0.3226*	1.0000				
novel	0.4454*	0.2848*	0.3009*	1.0000			
helpful	0.5240*	0.2840*	0.4262*	0.3651*	1.0000		
expert	0.4998*	0.2473*	0.3262*	0.4921*	0.4723*	1.0000	

Continuing to do data mining with this model to identify relative impact of individual aspects did not provide us any significant results. Any subset of these aspects could achieve a classification accuracy of about 80.00%, but this could simply be by chance given the nature of the data.

While the classifier built with this model was not highly successful, we did find the evidence of its correct functioning when we compared the probabilities computed for the instances of both the classes (‘yes’ and ‘no’). Table 3 shows the result of a two-tailed paired t -test between the probability distributions of ‘yes’ class and ‘no’ class. In the table, they are represented by ‘1’ and ‘0’ group respectively.

Table 3: Two-tailed paired t -test for probability distributions of ‘yes’ and ‘no’ classes.

Two-sample t test with equal variances

Group	obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	486	.1734148	.0049095	.108231	.1637684	.1830612
1	114	.2607054	.0123461	.1318206	.2362455	.2851653
combined	600	.19	.0048196	.1180562	.1805346	.1994654
diff		-.0872906	.0117663		-.1103988	-.0641824

diff = mean(0) - mean(1) t = -7.4187
 Ho: diff = 0 degrees of freedom = 598

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
 Pr(T < t) = 0.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 1.0000

As we can see, the mean probability for ‘no’ class instances was 0.1734, whereas for ‘yes’ class it was 0.2607. The test shows that the difference in these means was statistically significant. This indicates that even though the probabilities for class ‘yes’ were not higher than 0.5 (for their correct classification), they were significantly higher than those for class ‘no’.

In summary, we learned that it is extremely hard to pick the best answer from a set of answers for a given question. This is due to large variability in the subject and the kind of questions asked, the asker’s expectations, and the kind of answers given. In addition to this, only one out of many answers is selected as the best answer and it is very difficult to train a classifier using such training data, where not chosen answers may still be of good quality. We also realized that the 13 aspects of quality that we identified are highly correlated, but do not help us create a robust model for explaining the variability in the data, let alone predicting the best answer.

We did, however, find a high level of agreement among the workers while rating different aspects of answer quality (Figure 1). This informs us that while people have similar understanding or perception about the quality of an answer, their collective assessment was not enough to predict the decision of the asker.

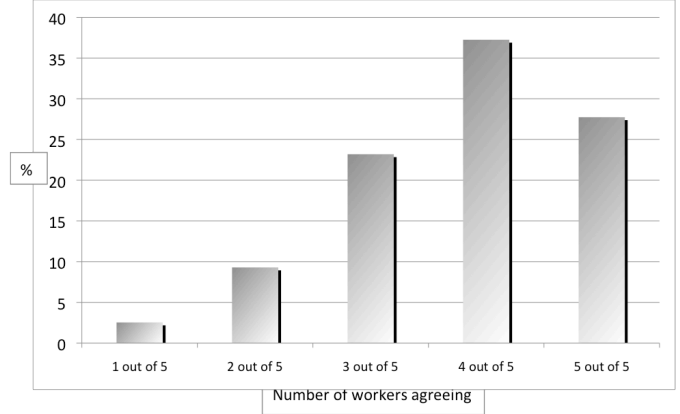


Figure 1: Agreement among 5 different workers while rating aspects of answer quality.

This led us to believe that there may be other aspects or features that we need to consider. In the next section we describe how we automatically extracted such features and used them for evaluating and predicting answer quality.

5. PREDICTION MODEL USING AUTOMATICALLY EXTRACTED FEATURES

Now we will describe a set of experiments done using automatically extracted features from questions and/or answers. As we discussed in the previous section, we may have a common perception of what constitutes to the quality of an answer, but in case of CQA, there are many other variables that may contribute to an asker choosing an answer to be the best answer for his information need. In particular, the profile of the asker as well as the answerer may play an important role in determining if the asker would find an answer appropriate for his posed question or not. In this section we will show how to use such user profile information to build a prediction model.

5.1 Extracting QA features

We extracted the following features for each question and answer in our dataset:

- Length of the question’s subject ($q_{subject}$)
- Length of the question’s content ($q_{content}$)
- Number of answers for the question ($num_{answers}$)
- Number of comments for the question ($num_{comments}$)
- Information from the asker’s profile (q_{points} , q_{level} , $q_{total_answers}$, $q_{best_answers}$, $q_{questions_asked}$, $q_{questions_resolved}$, q_{stars})
- Length of the answer’s content ($a_{content}$)
- Inclusion of references with the answer ($reference$)
- Reciprocal rank of the answer in the list of answers for the given question (a_{rr})
- Information from the answerer’s profile (a_{points} , a_{level} , $a_{total_answers}$, $a_{best_answers}$, $a_{questions_asked}$, $a_{questions_resolved}$, a_{stars})

A user’s profile (asker or answerer) contained the following information about that user: number of questions asked, number of those questions resolved, number of questions answered, number of those answers chosen as the best answers, level achieved in YA (1 to 7), number of points earned, and number of stars received.

Some of the data points had to be deleted because we could not obtain one or more of the features (e.g., a user’s profile taken down). This resulted in a total of 116 questions and 575 answers. Table 4 provides a summary of some of these features based on our dataset.

Table 4: Summary of various features used from questions, answers, and users (asker or answerer).

Feature	Min value	Max value	Mean
<i>Question</i>			
qsubject	12	110	54.92
qcontent	0	1030	195.46
numanswers	5	50	19.94
numcomments	0	3	0.15
<i>Answer</i>			
acomment	2	10625	182.77
reference	0	1	0.05
a_rr	0.02	1.00	0.35
<i>User (Asker/Answerer)</i>			
points	8	63785	521.35
level	1	7	1.19
stars	0	547	9.58
total_answers	0	16239	123.78
best_answers	0	2576	22.72
questions_asked	0	1071	26.47
questions_resolved	0	920	25.64

5.2 Constructing a model with logistic regression

Using these 21 features, we constructed a model with logistic regression; similar to what we did with human assessed 13 quality aspects reported earlier. This model is presented in Table 5.

As we can see, the model is quite good in terms of its power to explain the variability in the data. Since pseudo $R^2=0.1562$ with statistical significance, we could say that more than 15% of the variability in the data can be attributed to the factors used here.

The model was successful 84.17% times in selecting the best answer on the same training set, which was better than the prediction accuracy obtained from human assessments of quality factors. Doing 10-fold cross-validation resulted in 81.74% accuracy, which was once again better than the one achieved using human rated judgments.

Similar to the earlier model, we performed a two-tailed paired t -test between the probability distributions of the ‘yes’ or the ‘1’ class and the ‘no’ or the ‘0’ class. The test results are reported in Table 6. Once again, we find that even though the probabilities for class ‘yes’ are not higher than 0.5 on average, they are significantly higher than those of for ‘no’.

We also found that these different features are not all significantly correlated as we found earlier with the quality aspects assessed by

human workers. This indicates that not all the features extracted about questions and/or answers are geared toward predicting the quality of an answer.

Table 5: Logistic regression model for 21 automatically extracted QA features in predicting asker-rated quality of an answer.

rating	Coeff.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
qsubject	-.0038755	.0022726	-1.71	0.088	-.0083297	.0005787
qcontent	-.0007107	.0003948	-1.80	0.072	-.0014843	.0000632
numanswers	.017057	.0090116	1.89	0.058	.0006055	.0347195
numcomments	-.034717	.0907147	-0.38	0.702	-.2125146	.1430805
q_points	-.0022464	.0021258	-1.06	0.291	-.0064128	.00192
q_level	.1517189	.20077	0.76	0.450	-.241783	.5452208
q_total_an-s	.0048579	.0049749	0.98	0.329	-.0048927	.0146085
q_best_an-s	.0234524	.0215382	1.09	0.276	-.0187616	.0656664
q_quest1-ke	-.0354473	.0260461	-1.36	0.174	-.0864966	.0156021
q_quest1-ved	.0330828	.0242647	1.36	0.173	-.0144751	.0806407
q_stars	.0008161	.0049889	0.16	0.870	-.0089619	.0105941
acomment	.0020836	.0014807	1.41	0.159	.0008185	.0048857
reference	.2374713	.6453088	0.37	0.713	-1.027311	1.502253
a_points	.0053807	.0038277	1.41	0.160	-.0021215	.0128829
a_level	-.4164522	.3714064	-1.12	0.262	-1.144395	.3114909
a_total_an-s	-.0110888	.0085822	-1.29	0.196	-.0279096	.0057321
a_best_an-s	-.0594911	.0397994	-1.49	0.135	-.1374966	.0185143
a_quest1-ke	.1146048	.0913479	1.25	0.210	-.0644339	.2936435
a_quest1-ved	-.1071046	.0871774	-1.23	0.219	-.2779692	.06376
a_stars	-.0251896	.0180339	-1.40	0.162	-.0605353	.0101561
a_rr	.2906039	1.122684	0.26	0.793	-.5106464	1.7056184
_cons	-.9648778	.5064439	-1.91	0.057	-1.95749	.027734

Table 6: Two-tailed paired t -test for probability distributions of ‘yes’ and ‘no’ classes.

Group	obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	462	.156951	.0047909	.1029756	.1475364	.1663656
1	113	.3583066	.0224056	.2381746	.3139128	.4027003
combined	575	.1965217	.0067242	.1612397	.1833148	.2097287
diff		-.2013556	.0146999		-.2302279	-.1724832

diff = mean(0) - mean(1) t = -13.6977
 HO: diff = 0 degrees of freedom = 573

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
 Pr(T < t) = 0.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 1.0000

From Table 5, it appears that number of comments, existence of references with an answer, and the stars received by the asker are not helping much in the prediction. Once again, data mining with this model revealed that removing these three features retains the model’s ability to select the best answer without much loss (pseudo $R^2=0.1404$ and prediction accuracy=83.83%). Extending this further, we could get the most parsimonious model by having a_rr (reciprocal rank of an answer) only in the model (pseudo $R^2=0.0823$ and prediction accuracy=80.34%).

Not very surprisingly, features extracted from questions only do not help in prediction at all (Table 7, with significance reported to be 0.9594), whereas features extracted from answers only achieve quite high power (Table 8) with pseudo $R^2=0.1386$ and statistical significance. Once again, we find that the single most important feature that contributes to predicting if an answer would be selected as the best answer or not is the reciprocal rank of that answer (a_rr).

Finally, we performed likelihood-ratio test on the models constructed with 13 human-assessed aspects and 21 automatically extracted features. We found that both the models were significantly different ($\chi^2=17.29, p=0.0040$). This shows that the latter model was successful in outperforming the former one with statistical significance.

Table 7: Logistic regression model for 11 automatically extracted question features in predicting asker-rated quality of an answer.

```

Logistic regression      Number of obs =      575
                        Wald chi2(11) =      4.33
                        Prob > chi2 =      0.9594
Log pseudolikelihood = -284.91394      Pseudo R2 =      0.0001
    
```

(Std. Err. adjusted for 116 clusters in qid)

rating	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
qsubject	-.833e-06	.0009781	-0.01	0.993	-.0019253 .0019086
qcontent	.0000279	.0000349	0.80	0.423	-.0000404 .0000963
numanswers	-.0007592	.003645	-0.21	0.835	-.0079033 .0063848
numcomments	-.0209016	.0326137	-0.64	0.522	-.0848271 .043024
q_points	-.862e-06	.0000264	-0.33	0.744	-.0000603 .0000431
q_level	.0156938	.0094576	1.65	0.098	.0029016 .0342892
q_total_an-s	.0000205	.0000413	0.50	0.619	-.0000605 .0001016
q_best_ans-s	.0000674	.0004137	0.16	0.871	-.0007435 .0008784
q_questi-ke-d	-.0001519	.0006919	-0.22	0.826	-.001508 .0012042
q_questi-ved	.0001205	.0006348	0.19	0.849	-.0011238 .0013647
q_stars	-.0000234	.0000971	-0.24	0.810	-.0002136 .0001669
_cons	-1.403788	.1206554	-11.63	0.000	-1.640268 -1.167308

Table 8: Logistic regression model for 10 automatically extracted answer features in predicting asker-rated quality of an answer.

```

Logistic regression      Number of obs =      575
                        Wald chi2(10) =      444.51
                        Prob > chi2 =      0.0000
Log pseudolikelihood = -245.44934      Pseudo R2 =      0.1386
    
```

(Std. Err. adjusted for 116 clusters in qid)

rating	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
accontent	.0017766	.0012478	1.42	0.155	-.000669 .0042222
reference	.3128832	.6102943	0.51	0.608	-.8832717 1.5090338
a_points	.0031013	.0021472	1.44	0.149	-.0011071 .0073096
a_level	-.2672806	.2239349	-1.19	0.233	-.706185 .1716237
a_total_an-s	-.0062671	.0046254	-1.35	0.175	-.0153326 .0027985
a_best_ans-s	-.0350234	.0229574	-1.53	0.127	-.080019 .0099723
a_questi-ke-d	.0741423	.0724542	1.02	0.306	-.0678654 .2161499
a_questi-ved	-.0684709	.0692063	-0.99	0.322	-.2041127 .0671709
a_stars	-.0262627	.0208062	-1.26	0.207	-.067042 .0145166
a_rr	-2.859143	1.118054	-2.56	0.011	-5.050488 -.6677986
_cons	-.9574806	.40994	-2.34	0.020	-1.760948 -.1540131

5.3 Additional training and testing

Encouraged by the success of automatically extracted features for predicting answer quality, we extended our experiments to include another set of answers. Using YA APIs,¹¹ we randomly extracted 5032 answers from YA, associated with 1346 questions in different categories. We extracted all the 21 features reported earlier for each question-answer pair and constructed a logistic regression model. The model had pseudo $R^2=0.1312$, and gave 84.72% classification accuracy on the same data (Figure 2). When tried with 10-fold cross-validation, the model gave 84.52% accuracy, showing its robustness for classification.

When we used this model that we constructed with 5032 data-points for training and the previous model with 575 data-points for testing, the training model achieved 80.35% accuracy (Figure 3). This, once again, demonstrates the robustness of the selected features and the models constructed using them.

```

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances      4263      84.7178 %
Incorrectly Classified Instances    769      15.2822 %
Kappa statistic                    0.1808
Mean absolute error                 0.23
Root mean squared error             0.3365
Relative absolute error              83.2205 %
Root relative squared error         90.5292 %
Total Number of Instances          5032
    
```

Figure 2: Result of classification on the training data.

```

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances      462      80.3478 %
Incorrectly Classified Instances    113      19.6522 %
Kappa statistic                    0.0106
Mean absolute error                 0.2282
Root mean squared error             0.4168
Relative absolute error              76.8049 %
Root relative squared error         104.5803 %
Total Number of Instances          575
    
```

Figure 3: Result of classification on the test data.

All of the classification results reported in the present and the previous section are summarized in Table 9. It is clear that the models constructed with QA features (automatically extracted) tend to give reliable results on training as well as testing data, indicating their robustness.

Table 9: Summary of various classification models.

#	Model	Training	Testing	Classification accuracy
1	Quality aspects (human assessed)	600 samples	Same data	80.33%
2	Quality aspects (human assessed)	600 samples	10-fold cross-validation	79.50%
3	QA features (automatically extracted)	575 samples	Same data	84.17%
4	QA features (automatically extracted)	575 samples	10-fold cross-validation	81.74%
5	QA features (automatically extracted)	5032 samples	Same data	84.72%
6	QA features (automatically extracted)	5032 samples	10-fold cross-validation	84.52%
7	QA features (automatically extracted)	5032 samples	Data from model 3	80.35%

6. CONCLUSION

Measuring quality of content in community question-answering (CQA) sites presents a unique set of issues and challenges. As with many other IR problems, evaluating content quality is a significant challenge in CQA, and may require us to go beyond the traditional notion of “relevance” as suggested by Saracevic [11]. To address this challenge of evaluating answer quality, we used 13 different criteria to assess the overall quality of an answer on Yahoo! Answers (YA) site. We assumed that an answer is the best for a given question, if (1) the answer is chosen by the asker as the best, and (2) the asker gives it a rating of 3 or higher. This gave us a gold standard against which we could compare our models for evaluating and predicting answer quality. With the help of Amazon Mechanical Turk workers, we discovered that people have a good understanding and high level of agreement over what constitutes as a good quality answer. We also found that different aspects of the overall quality of an answer, such as informativeness, completeness, novelty, etc., are highly

¹¹ <http://developer.yahoo.com/answers/>

correlated. However, when these features were used for creating a model for predicting the best quality answers, we found it limiting. The human assessors were given the answers without any context (other than the question); they did not know who asked the question or who answered it, nor did they have any information about the conditions under which a question was asked or an answer provided. We realized that such information is critical in evaluating content quality in CQA.

We, therefore, extracted several features of the questions, the answers, and the users who provided them from YA. Via model building and classifying experiments, we discovered that indeed, the answerer's profile, as measured by the points earned on YA (social capital in that community), and the order of the answer in the list of answers for a given question, as measured by the reciprocal rank of that answer, are the most significant features for predicting the best quality answers. We demonstrated the robustness of our models by doing cross-validations and applying the trained models to a larger test set.

Beyond developing models to select best answers and evaluate the quality of answers, there are several important lessons to learn here for measuring content quality in CQA. It was pointed out, above, that there is huge variety in the kind of questions and answers found on CQA services, and that a given question may receive several answers from the community.¹² Given that only one of these answers can be picked by the asker as the best answer, it is extremely hard to create a classifier that outperforms random selection, based on *a priori* information about the data. However, with appropriate features, we could build models that can have significantly higher probability of identifying the best answer in the 'yes' class than of classifying a non-best answer in that class.

CQA provides unique opportunities to consider social factors and other contextual information, while evaluating its content. For instance, the placement of information, as well as the profile or social capital of its producer are some of the data that may help in better prediction and evaluation of content quality. A great many CQA sites exist, with different asking, answering, and rating mechanisms; and we believe that there are several other possible features of CQA questions and answers worth exploring. Future work will benefit from the unique issues presented here while evaluating and predicting content quality in CQA.

7. ACKNOWLEDGMENTS

We are grateful to anonymous workers of Amazon Mechanical Turk service for providing us valuable human assessments for the answers that we used here. We are also thankful to Yahoo! for making their QA datasets available to us.

8. REFERENCES

- [1] Dervin, B. (1998). Sense-making theory and practice: An overview of user interests in knowledge seeking and use. In *Journal of Knowledge Management*, 2(2), 36-46.
- [2] Gazan, R. (2008). Social annotations in digital library collections. *D-Lib Magazine*, 11/12(14). Available from <http://www.dlib.org/dlib/november08/gazan/11gazan.html>.
- [3] Harper, M. F., Raban, D. R., Rafaeli, S., & Konstan, J. K. (2008). Predictors of answer quality in online Q&A sites. In *Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems* (pp. 865-874). New York: ACM.
- [4] Janes, J. (2003). The Global Census of Digital Reference. In *5th Annual VRD Conference*. San Antonio, TX.
- [5] Kim, S., Oh, J-S., & Oh, S. (2007). Best-Answer Selection Criteria in a Social Q&A site from the User Oriented Relevance Perspective. *Proceeding of the 70th Annual Meeting of the American Society for Information Science and Technology (ASIST '07)*, 44.
- [6] Kresh, D. N. (2000). Offering High Quality Reference Service on the Web: The Collaborative Digital Reference Service (CDRS). *D-Lib Magazine*, 6.
- [7] Lee, J. H., Downie, J. S., & Cunningham, S. J. (2005). Challenges in cross-cultural/ multilingual music information seeking. In *Proceedings of the 6th International Society for Music Information Retrieval* (pp. 1-7). London, UK.
- [8] Liu, Y., Bian, J., & Agichtein, E. (2008). Predicting Information Seeker Satisfaction in Community Question Answering. *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*.
- [9] Pomerantz, J. (2008). Evaluation of Online Reference Services. *Bulletin of the American Society for Information Science and Technology*, 34(2), 15-19. Available from <http://www.asis.org/Bulletin/Dec-07/pomerantz.html>.
- [10] Pomerantz, J., Nicholson, S., Belanger, Y., & Lankes, R. D. (2004). The Current State of Digital Reference: Validation of a General Digital Reference Model through a Survey of Digital Reference Services. *Information Processing & Management*, 40(2), 347-363.
- [11] Saracevic, T. (1995). Evaluation of evaluation in information retrieval. *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval* (pp. 138-146). Seattle, USA.
- [12] Shah, C., Oh, J. S., & Oh, S. (2008). Exploring characteristics and effects of user participation in online social Q&A sites. *First Monday*, 13(9). Available from <http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fin/article/view/2182/2028>.
- [13] Shah, C., Oh, S., & Oh, J-S. (2009). Research Agenda for Social Q&A. *Library and Information Science Research*, 11(4), 205-209.
- [14] Su, Q., Pavlov, D., Chow, J., & Baker, W. (2007). Internet-scale collection of human-reviewed data. In C. L. Williamson, M. E. Zurko, P. E. Patel-Schneider, & P. J. Shenoy (Eds.), *Proceedings of the 16th International Conference on World Wide Web* (pp. 231-240). New York: ACM.
- [15] Voorhees, E. M (2003). Overview of the TREC 2003 question-answering track. In *TREC 2003*.
- [16] Zhu, Z., Bernhard, D., & Gurevych, I. (2009). A Multi-dimensional Model for Assessing the Quality of Answers in Social Q&A Sites. *Technical Report TUD-CS-2009-0158*. Technische Universität Darmstadt.

¹² Using more than a million questions from YA, we have found that on average a question receives about 6 answers.