What makes a high-quality user-generated answer?

Chua, Alton Yeow Kuan; Goh, Dion Hoe-Lian; John, B. M.

2011

John, B. M., Chua, A. Y., & Goh, D. H. (2011). What makes a high-quality user-generated answer? *IEEE Internet Computing*, 15(1), 66-71.

<https://hdl.handle.net/10356/100041>

<https://doi.org/10.1109/MIC.2011.23>

© 2011 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at: [<http://dx.doi.org/10.1109/MIC.2011.23>].

Downloaded on 26 Apr 2023 22:07:51 SGT

What Makes a High-Quality User-Generated Answer?

Blooma M. John, Alton Y.K. Chua, and Dion H.Goh

Abstract

Community-driven question-answering (CQA) services on the Internet let users share content in the form of questions and answers. Usually, questions attract multiple answers of varying quality from other users. A new approach aims to identify high-quality answers from candidate answers to questions that are semantically similar to the new question. Toward that end, the authors developed and tested a quality framework comprising social, textual, and content-appraisal features of user-generated answers in CQA services. Logistic-regression analysis revealed that content-appraisal features were the strongest predictor of quality. These features include dimensions such as comprehensiveness, truthfulness, and practicality.

Introduction

The Internet has radically altered the landscape for interaction. In the past, users could only visit static Web pages and access institutional content that mirrored print-based publications. Now they can easily create and distribute their own content. The shift from mere content consumption to content creation is fueled partly by the proliferation of social-computing applications such as community-driven question-answering (CQA) services. These services enable users to share content in the form of questions and answers. Examples of CQA services include Yahoo Answers (<http://answers.yahoo.com>) and Answerbag (www.answerbag.com).

Even though CQA services represent a viable channel through which users can meet their information needs, the lack of editorial control means that answers submitted for a question might be of varying quality. So, most CQA services allow other users to express their endorsement or disapproval of each candidate answer through a peer-rating system of stars or thumbs-up and thumbs-down counts. The question-answer cycle is closed when the asker manually selects the best answer. Figure 1 illustrates user activities in a typical CQA service.

Figure 1. A typical community-driven question-answering (CQA) service.

In essence, information quality assurance in current CQA services is driven mainly by the user community's goodwill. Furthermore, the makeup of a high-quality answer in CQA services remains largely a black box. So, we seek to uncover the attributes of the highest-quality answer from a group of candidate answers. To do this, we created a quality framework comprising different features of user-generated content in CQA services. Our study holds implications that might pave the way for

more robust CQA services that can identify the best answer, where available, without human intervention.

Related Work in Judging Information Quality

Recent studies have examined how people make judgments about information quality. Robert Taylor identifies the five dimensions of information quality as accuracy, completeness, currency, reliability, and validity [1]. Soo Young Rieh highlights source, content, format, presentation, currency, accuracy, and speed of loading as the bases for judging information quality [2]. Richard Wang and Diane Strong conceive quality along intrinsic, contextual, representational, and accessibility dimensions [3]. Carlo Batini and Monica Scannapieco define information quality as a multifaceted concept [4]. In particular, quality is highly influenced by the users' endorsement, even though this endorsement is subjective. These diverging concepts stem from information quality's amorphous nature.

We consider a high-quality answer to be the best available answer in the CQA corpus for a particular question. However, little research has specifically investigated quality regarding answers retrieved from a CQA corpus.

Research on identifying high-quality answers in CQA services has three major strands. The first uses social features—metrics that measure users' activities and interactions—to predict answer quality [5]. One such metric is click counts.

The second strand includes social features and textual features such as answer length, the number of unique words, and word overlap between the question and answer [6]. Textual features can be extracted directly from the answer's surface textual content. Studies on CQA services that use social and textual features have focused on answering factoid questions [7], predicting user satisfaction [8] and measuring the chances of obtaining a successful answer [9].

The third strand leans on qualitative comments from askers to determine the criteria for high-quality answers [10]. These criteria include the answers' correctness and comprehensiveness.

Research on information quality has also had a bearing on best-answer selection [11]. Informed by several recent studies on how people make judgments about information quality [12,13], we conceive answer selection as a process of relevance judgment. Four measures consistently associated with high information quality are accuracy, completeness, presentation, and reasonableness.

The Dataset

We used Yahoo Answers as our dataset. It's an exponentially growing resource for question answering, and many researchers use it [1]. Also, it's rich in not only content but also social features such as user ratings that we can incorporate in our quality framework. Finally, recent studies cited high user satisfaction with it and noted its knowledge-sharing potential [2]. In Yahoo Answers, askers manually select the best answer. So, in our study, we used best answers as a proxy for high-quality answers, insofar as they meet askers' information needs.

Our Quality Framework

On the basis of a review of related research, our framework comprises three categories of features (see Figure 2).

Figure 2. A quality framework for answer selection in CQA services.

Social features refer to the metrics that track users' participation and interaction in the CQA service: [3, 4, 5]

- Asker authority and answerer authority involve the number of points obtained for participation, including answering a question, voting for an answer, and having an answer selected as the best answer.
- Users' endorsement is the number of thumbs-up an answer attracts.

Textual features refer to metrics that can be culled from answers [2,4,5]:

- Answer length is the answer's word count.
- Question and answer length ratio is the ratio between the two lengths.
- Unique words is the number of unique non-stop words in the answer.
- Non-stop word overlap with the question is the number of identical non-stop words in both the question and the answer.
- High-frequency words is the number of words in the answer with a high corpus frequency.

Content-appraisal features refer to metrics assessing the answers' information quality [6, 7, 8]:

- Accuracy is the measure of the answer's correctness.
- Completeness is the extent to which the answer covers all relevant points.
- Presentation is the extent to which the answer is free from spelling and grammatical errors.

- Reasonableness is the extent to which an answer is deemed believable and internally consistent.

So, we measure 12 different features.

Feature Extraction and Analysis

We drew 400 questions from the Computers and Internet categories in Yahoo Answers. Each question attracted three answers, one of which the asker chose as the best answer. We wrote a Java program to extract all the social and textual features from our dataset. This is similar to techniques used in studies on quality estimation [4,5].

For content-appraisal features, we relied on experts' judgments as proxies for the users' judgments of the answers, in lieu of involving the original askers in this study. Other researchers have also used this approach [8,9]. For each question, we invited two evaluators knowledgeable in IT to rate all three answers on the basis of a five-point scale according to accuracy, completeness, presentation, and reasonableness. The evaluators also picked the best answer independently.

We obtained a Cohen kappa value of .74 between the two evaluators, indicating a nonchance level of agreement. To confirm the reliability of using the evaluators as proxies for the original askers, we obtained a Cohen kappa value of .86, which indicates a nonchance level of agreement between the evaluators' selection of the best answers and the original askers' selection. However, a limitation of our methodology is that we used only two evaluators. Nevertheless, the results serve as a preliminary indication of our approach's promise.

After we collated the data, we used logistic regression to test our framework.

Discussion of the Results

The framework's overall goodness of fit was statistically significant. We obtained an overall correct prediction rate of 89.3 percent. The logistic-regression analysis showed 10 significant features in identifying high-quality answers. On the basis of the variables' strength of association, we organized the features into four categories - strongly associated, weakly associated, negatively associated, and nonsignificant.

Table 1. Logistic-regression results.*

Feature category	Feature	β	Stand ard Error	Wald statistic	p	Exp(β)
Social	Asker's authority	.000	.000	3.879	.050	1.000
	Answerer's authority	.000	.000	2.205	.138	1.000
	Users' endorsement	.538	.184	8.547	.003y	1.712
Textual	Answer length	-.018	.007	5.854	.016y	.983
	Question and answer length ratio	.078	.030	6.702	.010y	1.081
	Unique words	.023	.010	5.052	.025y	1.023
	Non-stop word overlap with the question	-.066	.029	5.227	.022y	.936
	High-frequency words	.018	.007	6.360	.012y	1.019
Content appraisal	Accuracy	2.699	.352	58.833	.000y	14.866
	Completeness	3.036	.314	93.695	.000y	20.823
	Presentation	.672	.262	6.603	.010y	1.959
	Reasonableness	.330	.134	6.043	.014y	1.391

* β is the beta value, p is the significance, and Exp(β) is the odds ratio.

y p < .05

Features strongly associated with high-quality answers include completeness, accuracy, and users' endorsement. In other words, high-quality answers are invariably comprehensive and correct [4]. They're also intrinsically linked to the user's endorsement. Features weakly associated with high-quality answers are presentation, reasonableness, question and answer length ratio, unique words, and high-frequency words, in decreasing order of strength. Features negatively associated with high-quality answers include answer length and non-stop word overlap. In other words, lengthy answers and those resembling the questions linguistically might in fact turn out to be of poor quality. Finally, features with no bearing on high-quality answers are asker's authority and answerer's authority. As with the adage "never judge a book by its cover," neither the asker's nor the answerers' standing in the community is a predictor of high-quality answers.

Previous studies confirmed that the best answers in a CQA service are selected largely on the basis of their quality [4]. Our study extended existing research by examining answer quality at a more granular level. We found content-appraisal features, particularly completeness and accuracy, to be hallmarks of high-quality answers. Conversely, most social and textual features had a weak association or no association with answer quality.

Applying the Framework to CQA Services

With a deeper understanding of the makeup of user-generated answers' quality, the prospect of automated retrieval of best answers from a prepopulated CQA service could well become reality. We are developing a CQA system based on our quality framework; Figure 3 illustrates our system's architecture.

Figure 3. An automated CQA system.

We developed customized C# code that performs Web crawling to gather questions, answers, and related user profiles from Yahoo Answers. We performed the crawling on recently resolved questions in the four most popular categories. The initial crawling was limited to 10,000 questions, but the crawler updated regularly to expand the prepopulated CQA service.

The asker first posts a question through the prepopulated CQA service. The system maps the question to a set of similar questions that the system has already found, using a customized

clustering technique whose scope is outside this article. If the system finds no match, it returns no answers. Users will see a list of questions in the cluster that are most similar to the new question. Otherwise, a feature extractor captures significant social and textual features of the answers to the set of questions.

To compute each answer's quality index, we first assign heavy weights to completeness, accuracy, and users' endorsement. Then, we assign moderate weights to presentation, reasonableness, question and answer length ratio, unique words, and high-frequency words. Next, we assign negative weights to answer length and non-stop word overlap.

Given social and textual features' objective nature, the Java program we mentioned earlier can easily obtain their scores. For the automated computation of content-appraisal features, we adopt techniques used in previous studies [2,4]. In particular, to measure accuracy, we use Kullback-Leibler (KL) divergence, which is a probabilistic measure of difference, between an answer and the related information in Wikipedia. To measure completeness, we perform a pair-wise comparison of KL divergence between each answer and the other candidate answers. We measure presentation using a Flesh readability score that indicates comprehension difficulty. To determine the answer's reasonableness, we measure its consistency by comparing the most frequent domain-specific concepts used in the answers.

On the basis of the overall quality index computed for the set of features, the system returns high-quality answers to the asker. This eliminates the time the asker waits for the user community to respond, and hopefully paves the way for more robust CQA services.

To further establish our quality framework's validity, we could employ more evaluators from different backgrounds to rate our CQA dataset. Furthermore, we could expand the dataset to include content from other sources, such as Answerbag. This will let us test our model's generalizability across different CQA platforms and domains.

References

1. R.S. Taylor, *Value-Added Processes in Information Systems*, Ablex, 1986.
2. S.Y. Rieh, "Judgment of Information Quality and Cognitive Authority in the Web," *J. Am. Soc. for Information Science and Technology*, vol. 53, no. 2, 2002, pp. 145–161.
3. R.Y. Wang and D.M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *J. Management Information Systems*, vol. 12, no. 4, 1996, pp. 5–34.
4. C. Batini and M. Scannapieco, *Data Quality: Concepts, Methodologies and Techniques*, Springer, 2006.
5. J. Jeon et al., "A Framework to Predict the Quality of Answers with Non-textual Features," *Proc. 29th Ann. Conf. Special Interest Group in Information Retrieval (Sigir 06)*, ACM Press, 2006, pp. 228–235.
6. E. Agichtein, Y. Liu, and J. Bian, "Modeling Information Seeker Satisfaction in Community Question Answering," *ACM Trans. Knowledge Discovery from Data*, vol. 3, no. 2, 2009, pp. 1–27.
7. J. Bian et al., "Finding the Right Facts in the Crowd: Factoid Question Answering over Social Media," *Proc. 17th Int'l Conf. World Wide Web (WWW 08)*, ACM Press, 2008, pp. 467–476.
8. Y. Liu, J. Bian, and E. Agichtein, "Predicting Information Seeker Satisfaction in Community Question Answering," *Proc. 31st Ann. Int'l ACM Sigir Conf. Research and Development in Information Retrieval*, ACM Press, 2008, pp. 483–490.
9. Y. Liu and E. Agichtein, "You've Got Answers: Towards Personalized Models for Predicting Success in Community Question Answering," *Proc. 46th Ann. Meeting Assoc. for Computational Linguistics on Human Language Technologies: Short Papers*, Assoc. for Computational Linguistics, 2008, pp. 97–100.
10. S. Kim and S. Oh, "User's Relevance Criteria for Evaluating Answers in a Social Q&A Site," *J. Am. Soc. for Information Science and Technology*, vol. 60, no. 4, 2009, pp. 716–727.
11. M.J. Blooma, A.Y.K. Chua, and D.H. Goh, "A Predictive Framework for Retrieving the Best Answer," *Proc. 2008 ACM Symp. Applied Computing*, ACM Press, 2008, pp. 1107–1111.
12. T. Saracevic, "Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part III: Behavior and Effects of Relevance," *J. Am. Soc. for Information Science and Technology*, vol. 58, no. 13, 2007, pp. 2126–2144.

13. P. Katerattanakul and K. Siau, "Measuring Information Quality of Web Sites: Development of an Instrument," Proc. 20th Int'l Conf. Information Systems, Assoc. for Information Systems, 1999, pp. 279–285.