Evaluating Answer Quality across Knowledge Domains: Using Textual and Non-textual Features in Social Q&A

Hengyi Fu
School of Information Florida State University
hf13c@my.fsu.edu

Shuheng Wu
Graduate School of Library and Information Studies
Queens College, City University of New York
shuheng.wu@qc.cuny.edu

Sanghee Oh
School of Information Florida State University
shoh@cci.fsu.edu

ABSTRACT
As an increasing important source of information and knowledge, social questioning and answering sites (social Q&A) have attracted significant attention from industry and academia, as they address the challenge of evaluating and predicting the quality of answers on such sites. However, few previous studies examined the answer quality by considering knowledge domains or topics as a potential factor. To fill this gap, a model consisting of 24 textual and non-textual features of answers was developed in this study to evaluate and predict answer quality for social Q&A, and the model was applied to identify and compare useful features for predicting high-quality answers across four knowledge domains, including science, technology, art, and recreation. The findings indicate that review and user features are the most powerful indicators of high-quality answers regardless of knowledge domains, while the usefulness of textual features (length, structure, and writing style) varies across different knowledge domains. In the future, the findings could be applied to automatically assessing answer quality and quality control in social Q&A.

Keywords
Social question and answer sites, answer quality, quality assessment.

INTRODUCTION
Social questioning and answering sites (social Q&A) have become an increasingly important online platform for knowledge building and exchange. Social Q&A allows individuals to post questions, provide answers and comments, and evaluate questions and answers among the peers. The popularity of social Q&A has led to a growing body of literature regarding assessing and predicting answer quality. Most of the prior work has focused on this issue from two perspectives – users and data. The first group investigated what criteria users applied to evaluate answer quality (Kim & Oh, 2009; Shah & Pomerantz, 2010; Zhu, Bernhard, & Gurevych, 2009). Zhu et al. (2009) developed a multi-dimensional model for assessing answer quality from user perspectives, including informativeness, completeness, readability, conciseness, truthfulness, level of detail, originality, objectivity, novelty, usefulness, and expertise. This model was later used to evaluate answer qualities of Answerbag and Yahoo! Answers based on human judgement (Shah & Pomerantz, 2010; Zhu et al., 2009). The second group focused on extracting data features from questions and/or answers to automatically predict answer quality (Dalip, Gonçalves, Cristo, & Calado, 2013; Shah & Pomerantz, 2010), but the selection of features varies across studies and depends on the researchers’ perceptions of quality and the availability of different features.

Despite an ongoing concern pertaining to the answer quality in social Q&A, there is a lack of research and understanding of knowledge domain (topic) as a potential factor in assessing and predicting answer quality from both user and data perspectives. The criteria that a user applies to evaluate answers to science-related questions may be different from those of art-related questions. The usefulness of a particular quality criterion in predicting answer quality may also vary across knowledge domains. For example, the number of reference resources an answer includes may be highly correlated to the answer quality of science-related questions, but may not be very useful in predicting the answer quality of art-related questions. Harper, Raban, Rafaeli, and Konstan (2008) concluded that topics had a small and marginally significant effect on answer quality, but did not go further to study how it could influence quality assessment. Therefore, the purpose of this study is to fill this gap by (a) developing a model to evaluate and predict answer quality for social Q&A and (b) applying this model to identify and compare useful features of high-quality answers across different knowledge domains.

RELATED WORK
Answer quality in social Q&A can be captured either directly through textual features, such as content length or structure, or indirectly through non-textual features, such as review histories or author reputation. Textual features can be further divided into four groups: length, structure, style, and readability of texts. Non-textual features can be data
related to users, reviews, and networks in social Q&A (Dalip et al., 2013).

**Textual features**

*Length features* include the count of characters, words, or sentences, which give hints about whether the text is comprehensive. The intuition behind them is that a mature text of good quality should be neither too short, which may indicate an incomplete coverage, nor excessively long, which may indicate verbose contents. Length features have been proven to be one of the most powerful indicators of quality in both Wikipedia (Blumenstock, 2008; Hu et al., 2007) and social Q&A (Agichtein et al., 2008; Harper et al., 2008).

*Structure features* are indicators of how well the content of texts is organized. Previous studies on the quality of Wikipedia articles highlighted the significance of features that reflect organizational aspects of the article, such as the number of sections, images, external links, and citations (Hasan, André Gonçalves, Cristo, & Calado, 2009). Although the content structure of answers in social Q&A is not exactly the same as that of Wikipedia articles, some structure features can still be used to predict answer quality. For example, the presence of references and external resources are known as influential features of high-quality answers in social Q&A (Gazan, 2006).

*Style features* are intended to capture the author’s writing style. The intuition behind them is that high-quality content should present some distinguishable characteristics related to word usage, such as the use of auxiliary verbs, pronouns, conjunctions, prepositions, and short sentences. Style features have been proven as effective in predicting the quality of user-generated content in Wikipedia articles (Lipka & Stein, 2010) and answers in social Q&A (Hoang, Lee, Song, & Rim, 2008; Dalip et al., 2013), but the selection of style features varies among studies.

*Readability features* are intended to estimate the minimal age group necessary to comprehend a text. The intuition behind them is that high-quality content should be well written, understandable, and free from unnecessary complexity. Several measures of readability have been proposed, including the Gunning-Fog Index, the Flesch-Kincaid Formula, and SMOG Grading. These measures combine the number of syllables, words and sentences in the text, and were used to predict content quality in Wikipedia (Rassbach, Pincock, & Mingus, 2007) and social Q&A (Dalip et al., 2013), but resulted in low effectiveness in both settings.

Another set of textual features proposed specifically for online question answering is *relevance features*, identifying the similarity between the answer and its associated question, such as the number of words or sentences shared by the questions and answers. Dalip et al. (2013) tested relevance features using question-answer pairs from Stack Overflow, and concluded that these features were not very indicative of answer quality.

**Non-textual features**

*User features* reflect users’ activities and expertise levels in a social Q&A site. The intuition behind user features is that content quality can be inferred by examining the author who provides it. User features, such as account history (e.g., member since), community role, achievement, and reputation, have been used and proven to be significant quality predictors in different knowledge creation communities (Shah & Pomerantz, 2010; Stvilia, Smith, & Gasser, 2005; Tausczik & Pennebaker, 2011). In social Q&A, user features extracted from answerers’ profiles and the points that they earned were the most significant features for predicting the best quality answers (Shah & Pomerantz, 2010).

*Review features* are those extracted from the review history or comments on certain content. These features are useful for estimating the maturity and stability of content. It can be expected that a content receiving many edits or comments has likely been improved over time. In general, the more the content was reviewed, the better its quality. Positive correlation exists between content quality and review features, such as the number of edits/revisions, discussions, and comments (Lib, 2004; Stvilia et al., 2005; Wilkinson & Huberman, 2007).

*Network features* are those extracted from the connectivity network inherent within a community. The motivation for using these features to evaluate content quality is that citations between contents can provide evidence about their importance. For example, although a high-quality Wikipedia article should be stable, such stability may result from poor quality, and thus no one is interested in it. It is therefore important to take into account the popularity of content during quality evaluation, which can be estimated by measures taken from its connectivity network. Network features such as PageRank, link count, and translation count, have been tested in Dalip et al.’s (2011) study on Wikipedia article quality, but did not show high effectiveness.

**RESEARCH QUESTIONS**

This poster addresses the following research questions:

*RQ1*: What are the textual or non-textual features of social Q&A that can be useful to assess quality of answers?

*RQ2*: How do the features of high quality answers in social Q&A vary across different knowledge domains, especially in science, technology, art, and recreation?

**DATA COLLECTION AND RESEARCH METHOD**

To answer the first research question, this study selected a set of 24 quality features, which have been reported in literature review as useful for predicting the quality of contents in Wikipedia and social Q&A. Table 1 shows the
five feature groups of the 24 features. The usefulness of these quality features was tested and reported in this poster.

<table>
<thead>
<tr>
<th>Feature group</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>Counts of words (l-wc) and sentences (l-sc)</td>
</tr>
<tr>
<td>Structure</td>
<td>Counts of paragraphs (s_pc), images (s_ic), external links (s_elc), internal links (to other question/answers in the site) (s_ilc)</td>
</tr>
<tr>
<td>Style</td>
<td>Counts of auxiliary verbs (t_avc), “to-be” verb (t_bvc), pronoun (t_pc), conjunction (t_cc), interrogative pronoun (t_ipc), subordinating conjunction (t_scc), preposition (t_prp), nominalization (t_nc), and passive voice sentence (t_pvc)</td>
</tr>
<tr>
<td>User</td>
<td>Counts of user points (u_point), merit badges (u_mbc), questions posted by a user (u_qc), answers posted by a user (u_ac), comments posted by a user (u_cc), suggested edits (u_ec) made by a user</td>
</tr>
<tr>
<td>Review</td>
<td>Counts of revision (r_rc), and comments and answers received (r_cc), answer age/day (r_age)</td>
</tr>
</tbody>
</table>

Table 1. Quality features of five different groups

To answer the second research question, this study collected data from StackExchange, a group of popular social Q&A sites, currently with 133 different topical Stack Exchange websites hosting almost 3.8 million users. As of May 2015, 3.1 million questions have been posted, eliciting 4.5 million answers. The 133 social Q&A sites are grouped into six topic areas, including science, technology, business, professional, culture/recreation, and life/arts. This study selected four to represent different knowledge domains: science, technology, art, and recreation; and uses one site from each area based on their similarities in site age, membership, and active levels, namely Mathematics, Ask Ubuntu (Q&A for Ubuntu developers), Photography, and Arqade (Q&A for passionate video gamers), to compare the features associated with high quality answers in these sites.

High quality answers were identified using StackExchange’s Answer Score, which is calculated by the number of upvotes minus the number of downvotes that an answer received. With the help of StackExchange API, the top 1,000 answers with the highest scores were extracted to represent high-quality answers from each of the four sites. One of the researchers developed C++ code to collect data related to structure features of answer contents. Length and style features of answers were calculated by the GNU Style and Diction software program. Data related to user and review features associated with the 1,000 answers were extracted from the four sites directly using StackExchange API.

Exploratory factor analysis (EFA) was conducted to identify and compare the features in the five groups (factors) among the four sites using SPSS. EFA is a method aimed at extracting maximum variance from the dataset within each factor. For each dataset, all 24 features were tested using Maximum Likelihood as the extraction method and varimax rotation; factors having eigenvalues above 1 were retained. A cut-off of 0.6 was selected for analyzing factor loading since factor loading above this value is generally considered high (Field, 2005).

FINDINGS AND DISCUSSION

The results of EFA for each site are given in Table 2. Factor 1 for Mathematics contains two structure features: image count and external link count. Factor 2 is dominated by review features, including revision count and comment count. The count of comments that the author made is significantly related to this factor as well. All features in Factor 3 are related to user/author, including the count of merit badges, answers, comments, and suggested edits. Factor 4 is a combination of length (sentence count), structure (internal link count), and style (passive voice sentences count). This factor can be interpreted as content or text related.

For Ask Ubuntu, factor 1 is mainly about the user/author, including points, numbers of merit badges, comments, and suggested edits. Review features play an important role in factor 2. Similar to Mathematics, factor 3 has a good mix of length (sentence count), structure (external link count and internal link count), and style (preposition count).

Factor 1 for Photography contains three features: image count, revision count, and comment count. Factor 2 is mainly related to user features (merit badge count, answer count, and comment count a user made, and the answer age). Factor 3 includes two length features (counts of words and sentences) and one structure feature (external link count). Factor 4 is dominated by style features.

For Arqade, Factor 1 has three features related to the review (revision count and comment count) and user (user points) groups. Factor 2 contains only one length feature (word count). Factor 3 contains three features corresponding to features of the user group (merit badge count, answer count, and comment count a user made). Factor 4 is dominated by structure features (paragraph count, image count, and external link count). Factor 5 is mainly related to style features (pronoun count and subordinating conjunction count).

1.  http://stackexchange.com/sites?view=list#traffic
2.  https://api.stackexchange.com/docs/
The results show that the review and user features are the most useful groups, which are both non-textual features. Review features, especially revision count and comment count, are indicators of high-quality answers across the four sites. These features, measuring user engagement in an answer, could be the most effective indicators of answer quality. User features, such as the count of merit badges, answers, and comments a user made also present usefulness in identifying high-quality answers, regardless of the knowledge domains in the sites. This highlights the importance of user profiles and the history of a user in assessing answer quality.

The findings also indicate that the importance of textual features (length, structure, and writing style) vary across different knowledge domains. For example, structure features, such as image count and external link count, are the most essential features of high-quality answers in Mathematics Q&A, indicating user’s recognition of using equations (usually presented in image format) and references in science related answers. Similarly, image count is a powerful indicator of answer quality in Photography Q&A, suggesting that attaching real photos in answering these kinds of questions is essential. The length feature plays an important role in assessing answer quality in Photography and Arqade. This implies that for social Q&A addressing more “subjective” topics (e.g., art, recreation), people tend to consider that the answer length has a positive relation to answer quality. Style features, such as pronoun count, subordinating conjunction count, preposition count, nominalization count, and passive voice sentence count, are useful for identifying high-quality answers in art and recreation social Q&A sites, but not that important for those addressing science and technology questions. While review and user features are the most important feature groups across different knowledge domains, text features are also useful, and more importantly, less difficult to generate in terms of the preprocessing required.

**IMPLICATIONS AND CONCLUSION**

This study develops a model consisting of 24 quality features for assessing answer quality in social Q&A, and then uses the model to identify and compare useful features of high-quality answers in four different knowledge domains. This study has a few limitations in that StackExchange is one of the social Q&A sites; it may not represent all other social Q&A sites. Also, the four sites that have been randomly selected may not represent all other sites of each knowledge domain.

The proposed model of predicting answers with the textual and non-textual features from social Q&A can be applied for automatic assessment of answer quality and quality control in social Q&A. The model provides a comprehensive but critical set of quality features, and it can be used to build classifiers to select high-quality answers across knowledge domains. Future research can validate or expand this model by testing it in social Q&A sites from other knowledge domains. This study also reveals that high-quality answers in different knowledge domains are associated with a variety of quality features. Social Q&A designers can customize the reward and promotion system of different knowledge domains to encourage contributions that contain the most useful quality features for a particular domain. For example, since review features are strongly associated with quality across knowledge domains, more points could be awarded to those who submit comments or edit suggestions to answers provided by others. For science or technology, more points and merit badges can be granted to those who provide equations, source code, and references in their answers in social Q&A.

**REFERENCES**


