When the Crowd is Not Enough: Improving User Experience with Social Media through Automatic Quality Analysis

Dan Pelleg
Yahoo Labs, Israel
pellegd@acm.org

Oleg Rokhlenko
Yahoo Labs, Israel
olegro@yahoo-inc.com

Idan Szpektor
Yahoo Labs, Israel
idan@yahoo-inc.com

Eugene Agichtein
Emory University, USA
eugene@mathcs.emory.edu

Ido Guy
Yahoo Labs, Israel
idoguy@acm.org

ABSTRACT

Social media gives voice to the people, but also opens the door to low-quality contributions, which degrade the experience for the majority of users. To address the latter issue, the prevailing solution is to rely on the "wisdom of the crowds" to promote good content (e.g., via votes or "like" buttons), or to downgrade bad content. Unfortunately, such crowd feedback may be sparse, subjective, and slow to accumulate. In this paper, we investigate the effects, on the users, of automatically filtering question-answering content, using a combination of syntactic, semantic, and social signals. Using this filtering, a large-scale experiment with real users was performed to measure the resulting engagement and satisfaction. To our knowledge, this experiment represents the first reported large-scale user study of automatically curating social media content in real time. Our results show that automated quality filtering indeed improves user engagement, usually aligning with, and often outperforming, crowd-based quality judgments.

Author Keywords: Automatic quality evaluation; Quantitative analysis; A/B testing; User engagement

ACM Classification Keywords: H.5.3 [Group and Organization Interface]: Web-based interaction

INTRODUCTION

The Internet is famous for lowering the barrier of entry to content publishing, and nowhere has it been more clear than on sites based on social and user-generated content (UGC). In contrast to the model where content is authored by a select few and curated by even fewer, the UGC model allows anyone to author content, and often to also express an opinion on other users' contributions. The downside of this democratization is the proliferation of low-quality content, be it from careless or lazy authors, from spammers, trolls, or sometimes even from users confused by a site's user interface. This issue is brought to the forefront by web search engines, which aggressively index UGC content, such as forums and community question-answering (CQA) sites, and often surface this content to the searchers. Thus, the content posted in UGC sites achieves a far wider audience than merely the original users of the site, and content quality filtering becomes especially critical.

The most common solution to the quality problem is letting the crowd filter the content. After all, this is the resource that scales up with content generators. Sites use a variety of user rating mechanisms, with the most popular being "thumbs up" and "thumbs down". Typically, a voting model enables users to endorse (up-vote), and sometimes vote against (down-vote) pieces of content. This is complemented by a common user interface for viewing the content ranked by votes already given, promoting the content liked best by the "crowd". However, as we show in this paper, the most popular, or the most highly voted content by the crowd, is not necessarily the best one to show to other users. One issue is the subjectivity of the votes, as we discuss later. Another issue is voting sparsity — many posts will never be voted on, while others will be voted on sparsely and for reasons that do not align with the needs of users searching for information relating to this content. Instead, algorithmic approaches to quality assessment of UGC have been proposed, which use a combination of social, semantic, and syntactic signals to score contributions in CQA, and similar user generated content sites (e.g., [1, 21, 37, 19]).

Automatic evaluation of quality has been shown highly effective according to a variety of manual assessments, ranging from expert editors [1] to MTurk workers [37]. However, it is not clear what the effect is on the actual users of the UGC site. Do users prefer to see all the content, including poor-quality contributions? And, in the context of information seeking sites like CQA, does filtering out poor content improve users' satisfaction? As far as we know, there has been no reports in the literature on how automated content curation affects user satisfaction and engagement within a UGC site.

The main focus of this work is on exploring the *effects* of deploying automated quality estimation system in a live production with millions of users. We study this by instrumenting and logging live user traffic on a major community question answering site, complemented with extensive manual assess-

CSCW'16, February 27-March 02, 2016, San Francisco, CA, USA
 ISBN 978-1-4503-3592-8/16/02...$15.00
DOI: http://dx.doi.org/10.1145/2818048.2820022

ments of the actual posted content and the user responses. Our main findings show that algorithmic quality scoring:

- Improves the accuracy of content recommendation, compared to filtering based on crowd votes alone.
- Promotes longer dwell times, as well as deeper exploration of content.
- Outperforms the asker-chosen selection of a best answer.
- Outperforms the voter-chosen selection of a best answer in 99% of the corpus.

Our system has been deployed on Yahoo Answers, one of the leading CQA sites, and has been powering the default display order of answers. To our knowledge, this is the first time that user-generated content is being curated algorithmically within a live system, experienced by millions of users.

The contributions of this paper are:

1. First reported study of deployment of an automatic system to evaluate the quality of user submissions to a leading CQA site.
2. First empirical proof that use of algorithmic quality judgment increases user engagement.
3. Extensive evaluation of the algorithm against objective human judgement, showing it outperforms the asker's opinion.
4. Evaluation against the crowd's rating, showing that our algorithm performs better than crowds of up to 20 users.

In combination, these contributions establish the benefits of algorithmic content curation of user generated content in a major community question answering site, with implications extending to other forms of crowd-based content curation.

RELATED WORK

This work spans two main areas: 1) analyzing user generated content for automatic quality evaluation, and 2) measurement of user engagement and satisfaction in social media.

Content Quality Curation in UGC

The problem of automated quality estimation of content on the web has been an active area of research for over a decade. Most solutions filter the content based on feedback from the crowd. Sites use a variety of user rating mechanisms, with the most popular being "thumbs up" and "thumbs down" (or up- and down-votes), available in most if not all UGC or CQA sites. However, as we show in this paper, the most highly voted content by the crowd is not necessarily the best one to show to other users. The sources of the discrepancy is the subjectivity in user voting, as well as other confounding factors. For example, if a user deeply agrees with the opinion or sentiment expressed, a positive vote is likely even if the content itself is worded poorly or illogical (and conversely, disagreement with an eloquently-expressed opinion could imply a negative vote). This causes bias towards extreme and emotionally provoking content [20]. Other reasons include voting by "troll" users, and voting wars among other users. Moreover, because voters are exposed to previous votes, phenomena such as the "rich get richer" [35] and "social influence bias" [32] diminish the signal from all but very few of the earlier posts.

Instead, algorithmic approaches to quality assessment of UGC have been proposed, which use a combination of social, semantic, and syntactic signals to score contributions. Indeed, while social media exhibits a number of specialized characteristics, the basic ideas of textual content analysis predate UGC. In the field of Automated Essay Scoring (AES), writings of students are graded by machines on several aspects, including writing style, information accuracy, and structure. Another approach is studying text quality from a readability perspective, and several measures have been proposed, such as the Flesch-Kincaid formula [24], which has been extended since for the Web setting (e.g., [22, 39]).

In addition to textual content, link-based methods have been successful for several tasks on the Web and in social media. In particular, link-based ranking algorithms were successful in estimating the quality of web pages, two of the most prominent being PageRank [33] and HITS [25]. Interestingly, applying link analysis effectively is more challenging in social media than on the Web, partly because the barrier to entry of creating a link is much lower [1]; however, in combination with text content analysis, link analysis has been shown to improve quality estimation [1, 21, 19] through direct feature augmentation or reinforcement [4]. Reputation and other metadata also turn out to be useful in estimating content credibility on Twitter [5]. In another active area of research, there are efforts in estimating product review quality. Those integrate text features, reputation, links, and explicit ratings from other users (e.g., ratings of review helpfulness) [8] or the price premium that good reviews can offer [12]).

In the community question-answering setting, it is less clear how to apply link-based and other meta data: links and community ratings require a long time to accumulate, but the content contributors and readers require feedback within minutes of posting the content. The most commonly used proxy is author reputation, as this signal is available immediately and can be an effective indicator of post quality. Most state of the art answer quality scoring (AQS) systems use a combination of content features, link analysis, and community feedback. For example, one of the earlier works [21] extracted a set of features from a sample of answers in Naver, a Korean question-answering portal, which derived answer quality based on features of the answer text, explicit user feedback, and user reputation. This work has been substantially extended in the years since with additional features and analysis [1, 16, 37, 19, 2, 36]. The AQS system we developed is similar in spirit, but uses a wider range of features including both structural, textual, and reputation, while remaining fast and practical for use at large scale.

User Engagement in Community Systems

One of the biggest factors in improving the user experience with web services has been extensive instrumentation to track and understand user behavior and engagement [27]. We build on previous work in user behavior modeling, particularly web search behavior, which has studied in depth the signals and metrics for that data. Among these, clicks on results and page

dwell time [28] (the time a user spends on a page) have been shown to be effective for result relevance prediction [13] and searcher satisfaction estimation [17]. These efforts have been extended to take advantage of additional user behavior signals such as cursor movements and scrolling [13]. With the emergence of mobile devices and touch screens, the variety of signals available has become even richer, resulting in more accurate models of searcher satisfaction [14] and result relevance [26]. One particularly important direction has been the study of a specific subset of user behavior analysis focusing on *user engagement* [27], indicative not only of short-term satisfaction but also long-term happiness and return rates to the site [9, 10]. While previous work demonstrated that it is possible to infer page quality and relevance from behavior, our work for the first time investigates the reverse direction: how does content quality *affect* user behavior. Until this work, it remained unknown if editorial ratings, whether from experts or crowdsourced, agree with the behavior of users who actually benefit from the content. Our work thus fills an important gap between theory and practice by demonstrating how the users react to deployment of a state of the art AQS system.

ALGORITHMIC QUALITY SCORING

In recent years, a significant amount of research was conducted on the task of answer quality assessment [21, 1, 4, 41, 37, 34, 42, 7, 19]. Our AQS algorithm largely follows the insights drawn by this body of prior work. Still, some of our decisions may deviate from the traditional view of the task and some of the features we employ are not common. Therefore, in this section we formalize the AQS task and detail our algorithm as we implemented it.

Task Definition

Like most prior work [1, 4, 37, 19], we think that an answer has an absolute notion of quality, and a human observing it can decide whether it is a high or low quality answer for the corresponding question. We therefore formalize the task as a classification problem; the input is an answer and its *context*, which includes the answer's content, the answer's meta data, the question's contents, and other answers posted in response to it. The output of the algorithm is a binary prediction whether the answer is of high or low quality.

Some works view user rating as a proxy for high-quality content, either looking at best answers [4, 37] or user voting (thumbs up or down) [7]. Yet, we follow a different perspective, which states that user rating by itself may not be sufficient to differentiate between high-quality and low-quality content [1, 34, 38]. Specifically, best answers may not be the only high quality answers, and voting with thumbs up or down may also capture a notion of viewer agreement and not necessarily of quality. We did not ignore the valuable information provided by user ratings and utilized them as signals within our classification scheme (see below). Yet, as an absolute source of truth, we asked human annotators for their assessment of high and low quality. These annotations served as a gold standard for training and testing our algorithm.

Next, we detail the various features that we extracted for our algorithm, followed by a description of our classification scheme and its performance on a manually annotated test-set.

Quality-related Features

Most of the features we induced from the input (answer and context) had appeared in some form in prior work. We employed a large and varied set of signals, and for the sake of completeness and reproducibility, we detail the various feature families below. We provide references for feature families that were found effective in prior work and further motivate novel features that we introduce.

Text Style This feature family measures the writing style of the answerer in terms of her word selection [1, 7]. These features include counting (separately) the number of misspellings, stop words, abusive words, polite words, articles, pronouns and prepositions in the answer's text. In addition, we also counted phrases that are common as short answers, which indicate an answer bearing empty content. Some examples are "*yes*", "*no*", "*idk*", "*sure*" and "*i think so*". For each respective feature, we created representations as a raw count, as a ratio, and as a binary indicator (nonzero appearance count).

Text Statistics Another notion of answering style concerns overall text statistics [7, 19]. We measured the answer length (both character and word counts), the average word length, the percentage of punctuation marks, and the percentage of capitalized sentences and capitalized words. We also counted the number of hyperlinks in the answer.

Best Answer Language Model Prior work utilized language models as a proxy for grammaticality and quality [1, 7]. Similarly, we constructed a trigram language model over 1 million best answers (all chosen by the respective askers) as a corpus of expected style of relevant answers (as viewed by the askers). We induced a feature of the log likelihood of the target answer text, under this language model.

User Feedback These features include the user feedback given to the answer on site [4]. These are indicator features for best-answer, the number of thumbs up and thumbs down given to it, and the number of edits the answerer performed on her answer.

Answerer Reputation These features capture aspects of the answerer's reputation, which may indicate her ability to generate high quality answers [37, 7]. For the author of the answer, we generated the number of thumbs up and thumbs down on past answers, best answer absolute number and ratio, the number of comments the past answers received, the total number of answers, best answers, questions, comments, votes, stars and thumbs provided by the answerer, and her overall tenure on the site. In addition, we tracked the number of points for the user (Yahoo Answers awards points for various user actions).

Surface Word Question Similarity We expect that the similarity of the answer text to the question indicates its relevance to the question and is thus a signal for higher quality. As a basic textual similarity, we computed the cosine similarity

between the word vector of the answer and the question. The weight of each word is its tf·idf score, and stems instead of words were maintained (using Lucene's Porter stemmer). We computed two similarity values for the answer: one compared it just to the question title, and the other compared to the concatenation of the question title and question body.

In addition, we used the following four features, which are novel for answer quality scoring, to the best of our knowledge.

ESA-based Question Similarity Since question and answer "languages" differ, one may expect low surface word similarity between questions and answers, even for relevant ones. To overcome this difference in wording, we represented each text by its Explicit Semantic Analysis (ESA) vector, within the space of Wikipedia's concepts [11]. We then computed the cosine similarity between the vectors of the answer and the question.

Answer Similarity We expect that repeated recommendations or opinions in different answers would indicate that they represent more important relevant information or a more common view. Therefore, answers with information that appears in other answers may be of higher quality. We therefore computed the average textual similarity between each target answer and the other answers. We computed both the surface-word average similarity and the ESA-based average similarity.

Query Performance Predictor The focus of an answer on a specific informative topic may be a good indicator that the answer provides useful and valuable information to the asker. We measured how much a text is "focused" by computing the Clarity [6] and Query Feedback [43] query performance predictors of the answer. These predictors analyze the resulting documents returned for the answer when issued as a query to the Lucene[1] search engine, over an index of a random sample of 2 million question/best-answer pairs. At a high level, these measures look at the difference between the language model defined by the retrieved documents and the language model of the general corpus (the 2 million question/best-answer documents). The more the retrieved language model differs from the general corpus, the more focused it is assumed to be.

Sentiment Analysis Putting the actual content aside and looking just at the wording, empathic answers are appealing, while "flaming" text in an answer alienates the reader. To capture this intuition, we utilized the SentiStrength[2] tool to extract the positive, negative, and neutral sentiment levels of the answer. While sentiment analysis has been recently used to classify speech acts in forums posts within a MOOC [3], to the best of our knowledge its use for AQS is novel.

Learning and Offline Evaluation

As stated above, we think that the notion of high or low quality is absolute and that more than one answer can be of high quality, and on the other hand, there may be no high-quality answer provided. To construct a gold standard data-set for training and testing, we used 40 professional human annotators (raters)[3]. The annotators were asked to choose whether an answer is '*Excellent*', '*Good*', '*Fair*', or '*Bad*'. Inspecting some of the annotations, we found that annotators would often select '*Fair*' to indicate that they were not completely sure about the exact quality of the answer. Specifically, raters would rarely mix between '*Good*' and '*Bad*' annotations, but would sometimes mark a '*Good*' answer as '*Fair*', and similarly a '*Bad*' as '*Fair*'. The bin for '*Fair*' therefore became too noisy. To reduce this noise, we decided to ignore all '*Fair*' annotations and selected '*Excellent*' and '*Good*' as high-quality examples, and '*Bad*' as low-quality examples.

Altogether, the above annotation process, after removing the '*Fair*' examples, resulted in 12,373 labeled answers for 6,038 questions. Of these answers, 65.2% were labeled as high quality and the rest as low quality. We split the questions into training and test sets, 11,114 and 1,259 in size, respectively (we split by questions, taking all labeled answers per question either into the training set or the test set). We compared several standard classification algorithms using Weka[4], and chose the best-performing logistic regression as our classifier. We measured the performance of our algorithm over the test-set by area under the ROC curve (AUC), achieving an AUC value of 0.81. This regressor also has the useful property of generating a confidence value (the likelihood of a high-quality classification). We made use of this property above, to rank the answers by decreasing quality.

The most significant features, in addition to user feedback, were a mix of answerer reputation, answer text, and question-answer similarity. More specifically, they included the answerer's absolute number of best answers; the answerer's tenure on the site (longer tenure was indicative of higher quality); the punctuation percentage in the answer text (more punctuation was indicative of higher quality, as it indicates a higher effort from the answerer); the match to the best answer language model; and the ESA similarity between the answer and the question. A pairwise cross-correlation analysis among the top and novel features indicated there is no high correlation (above Pearson's $r = 0.4$) between any pair of features. Moderate correlation was found between the ESA-based question similarity and answer similarity ($r = 0.39$, $p < .001$); the answerer's total number of best answers and the answerer's tenure on the site ($r = 0.23$, $p < .001$); and the ESA-based question similarity and the query performance predictor ($r = 0.22$, $p < .001$). None of the features was correlated with the number of positive or negative thumbs. Overall, the features appear to be complementary and no obvious redundancy could be observed.

In the following two sections, we describe our experiments, which consisted of two main parts. In the first, we examined the effectiveness of the AQS algorithm described in this section, by comparing its best scored answer with the best answer selected by the question asker and by the answer's voters. In the second part, we examined the effect of our algorithm on user engagement, by considering the relative po-

[1] http://lucene.apache.org
[2] http://sentistrength.wlv.ac.uk/
[3] In-house, rather than crowdsourced.
[4] http://www.cs.waikato.ac.nz/ml/weka

sition of clicked-answers, as well as dwell-time and scrolling depth on the question page.

EVALUATION OF QUALITY

Our first set of experiments examined the quality of the AQS algorithm by comparing its best ranked answer to the best answer as determined by two groups of users, each with their own respective degree of interest, attention span, and knowledge. The first group included the askers, who have a declared interest strong enough to have posted the question in the first place, but possibly little knowledge (or else they would not have asked). The second group of users included the site visitors, more specifically those who voted for answers. Their interest might be more fleeting, but on the other hand they felt knowledgeable enough to pass judgment. This second group is what is typically called the "crowd", whose wisdom is commonly referred to for evaluating user-generated content. Both experiments were performed on a massive scale, over a huge corpus of questions. We augmented these with human judgments, this time on a smaller scale, to reconcile the ties.

Best Answer by Asker vs. Best Answer by Quality

In the first experiment to assess the performance of our AQS algorithm, we measured the congruence between the algorithmic best quality answer, and the best answer as chosen by the asker. The asker's choice has been extensively used in the literature as the gold standard for answer quality, as the asker is expected to be the best judge of whether an answer satisfies their needs (e.g., [4, 30, 37]).

We analyzed a set of over 100 million questions from Yahoo Answers, posted between 2005 and 2014. Of these, 34% contained best answer chosen by the asker. In 63% of the cases, the best answer by algorithmic quality was the same as the best answer chosen by the asker. Out of the remaining 37% of the cases, we uniformly sampled 500 questions and performed an editorial study. In this study, the raters were presented with a question and two answers: the best answer chosen by asker and the best answer by the AQS algorithm, and were asked to decide whether one of the answers is better. Specifically, they were instructed to *"read the question and both answers, decide which of the two answers is a better answer to the question, and check the corresponding checkbox"*, where the available check-boxes were: *"Answer 1"*, *"Answer 2"*, *"Both are good"*, and *"Both are bad"*.

Figure 1 summarizes the results. In 46% of cases, both answers were of equal quality (40% both good and 6% both bad). AQS won in 37% of the cases (68% of non-tie cases) and best answer by asker – in 17%. This difference is statistically significant at $p < 0.05$ (Wilcoxon double-sided signed-rank test).

Best Answer by "Crowd" vs. Best Answer by Quality

In our second experiment, we wanted to measure the congruence of the best quality answer by the AQS algorithm and the best answer judged by user feedback. Users provide their feedback in Yahoo Answers by thumbing up or down a specific answer. We define herein the user feedback value on an answer as the number of thumbs up minus the number of thumbs down.

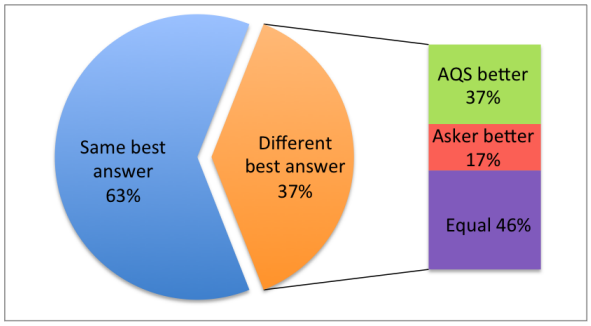

Figure 1. Best answer by asker vs. AQS. The pie on the left indicates the portion of identical versus different best answers over a huge corpus of questions. The bar on the right shows the comparison between the two when answers were different, based on an editorial study over a small sample.

We inspected again the dataset of over 100 million questions from Yahoo Answers mentioned in the previous sub-section. Nearly 92% of the questions had at least one user feedback (either up or down), but for over 99% of the questions, there were no more than 20 thumbs in total. Overall, in 29% of the questions, the best algorithmic quality answer agreed with the best answer by user feedback. This is a substantially lower portion than the agreement with the best answer by asker. For the remaining 71%, we set out to compare the two methods and also examine the effect of total number of thumbs on this comparison. We therefore sampled 500 questions out of this portion of 71%, stratifying by the number of thumbs, 100 questions for each of the following strata: up to 5 thumbs, 6–20 thumbs, 21–50 thumbs, 51–100 thumbs, and over 100 thumbs. While this stratification is far from representing the entire question dataset, it allowed us to closely inspect the influence of high number of votes. For each question, we extracted the answer with the best user feedback and the best algorithmic quality answer. Then we asked human raters to indicate if one of the two is better, or if both answers are of the same quality, as done in the previous experiment.

In 41% of cases, the answers were indicated to be of equal quality. In 30% of the cases (51% of non-ties), the best answer by AQS was chosen as better and in 29%, the best answer by user feedback was chosen as better, however this difference was not statistically significant.

On the other hand, considering only questions with 20 thumbs or less (which, as mentioned, comprise over 99% of all questions in the corpus), the AQS best answer was significantly better than the user-rated one, chosen in 36% of the cases (57% of non-tie cases), compared to 27%, respectively ($p < 0.05$, Wilcoxon double-sided signed-rank test). Figure 2 summarizes these results. For questions with more than 20 answers (less than 1% of the dataset), there was a significant difference in favor of user feedback-based best answers, at 30% vs. 26% ($p < 0.05$, Wilcoxon double-sided signed-rank test).

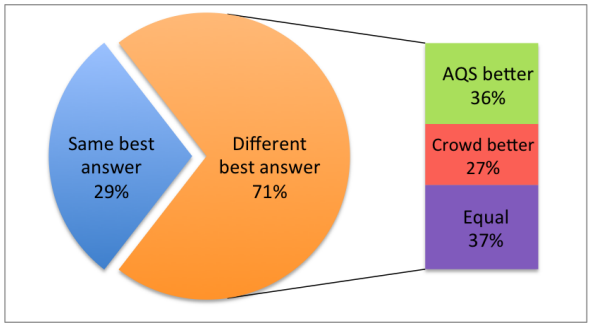

Figure 2. Best answer by "crowd" vs. AQS. The pie on the left indicates the portion of identical versus different best answers over a huge corpus of questions. The bar on the right shows the comparison between the two when answers were different, based on an editorial study over a small non-stratified sample.

EFFECT ON ENGAGEMENT

After validating the quality of our AQS algorithm, we set out to explore whether the algorithm serves the desired outcome when exposed to actual site users. Our second set of experiments therefore focuses on in-vivo tests, inspecting whether our method is beneficial in terms of user engagement.

Position of Clicked Answers

Our first user-engagement experiment focused on clicked answer position and was performed using A/B testing on live user traffic. *A/B testing* (sometimes referred to as "split testing" or "bucket testing") is an evaluation method that compares two variants, the "control" and the "treatment", through a controlled experiment, in which some users receive the control variant and others receive the treatment variant. It is currently the industry standard for evaluating website features on a large scale. In our case, for the control variant, we ordered the answers for each question using user thumbs, promoting answers that had the biggest difference between the number of thumbs-up and thumbs-down. For the treatment version, we ordered the answers for each question using algorithmic quality. In addition, we hid the answers with score lower than a threshold, set to the top quality score minus a parameter called α. We then measured both variants using a specially instrumented version of the Yahoo Answers landing pages, described below.

Instrumentation

The standard Yahoo Answers user interface (UI) includes a question page, where all of the answers are visible, but does not include any user controls that could be instrumented (e.g., clicks). We therefore modified the user interface by truncating each answer text after 2 lines, and adding a teaser link labeled "show more". Clicking on the teaser link exposed the rest of the answer in-line. Figure 3 illustrates the modified UI. The clickthrough-based metrics described below refer to the click needed to expand the answer.

Metrics

Our metrics were computed based on clicks on the "show more" link, which provided a fine-grained measure of interest in each answer. As a basic metric we used the Click-Through

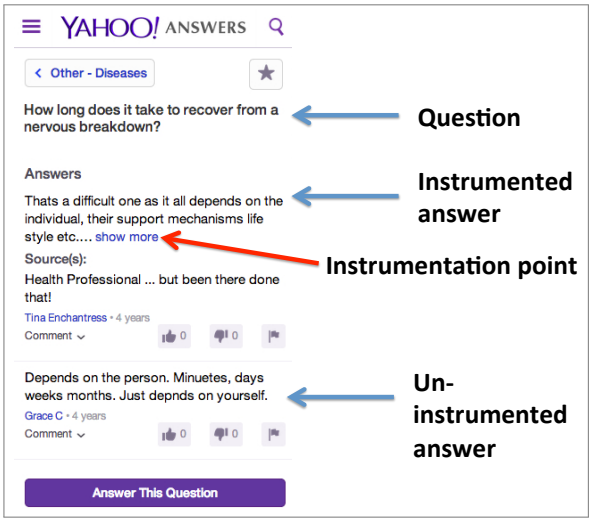

Figure 3. Instrumentation of content display in the modified Yahoo Answers user interface. The first answer is longer than 2 lines and is therefore truncated. The "Show more" teaser link is added and instrumented to require an explicit click to view the hidden content. The second answer is too short to be instrumented.

Rate (CTR) – a standard way to measure the level of interest in the presented content. In our setting, the CTR is measured as the ratio between the number of clicks on the "show more" link and the number of times it was presented. Statistically, one can view the impression-click relationship as a binomial process where the click-through rate reflects the probability p of a success (click) in a trial (impression). The maximum-likelihood estimate of p is then simply the number of observed successes (clicks on the "show more" link), divided by the number of trials, i.e., the number of times the link was shown. Higher values of CTR suggest higher user engagement.

To augment CTR, we used Mean Reciprocal Rank (MRR), which measures how high in the list the click occurred (i.e., how highly ranked was the first answer that was expanded). More formally, MRR is defined as a multiplicative inverse of the rank of the first relevant result. The higher it is, the better, with the best case being MRR=1 (when the chosen result is at the top slot) and the worst case being MRR=0 (when no click occurs). MRR is commonly used in information retrieval for evaluating any process that produces a list of possible responses to a query [31]. In our case, the query is a posted question, the responses are the answers, and the first relevant result is the first clicked answer. Therefore, the more successful ranking of the answers would result in more clicks on the top answers and therefore a higher MRR.

Experimental setting

The experiment was performed on live mobile user traffic over a period of two weeks. During this period, we collected hundreds of thousands of page views. In the default mobile Yahoo Answers UI, the question is presented on the top, followed by the best answer, and then the remaining answers, ordered by user feedback. The question page can contain a maximum of 5 answers and in order to see the other answers,

a user needs to click the "next page" button. To perform our experiment, we tweaked the standard UI by (1) disabling the reserved slot for the best answer, and (2) truncating all the answers to a maximum of two lines with a teaser link, as previously explained. Answers too short to contain a teaser link (19% of all answers) were not counted as an impression and therefore excluded from the measurement. Also, we disregarded impressions and clicks below position 5 as only a small percentage of the users used the "Next page" link.

Obviously, short answers, along with low-quality hidden answers, change the number of alternatives for a user to click on and therefore affect the metrics. In order to ensure a fair comparison between control and treatment, we binned all question page views by the number of answers available for click, and computed the CTR and MRR for every bin separately. Finally, both metrics were aggregated across bins.

Results

In all our experiments we used $\alpha = 0.5$. Our analysis show that the treatment variant outperformed the control variant by 9.2% in terms of CTR and by 3.8% in terms of MRR. Both results are statistically significant with $p < 0.01$ using Hoeffding's bound [18].

We also tested a simplified version of the treatment, which only ranked the answers by their quality score but did not hide the low-quality ones. The performance of this variant was more modest with a 5.5% increase in CTR and a 2.8% increase in MRR, as compared to the control[5].

User Engagement in Exploring Answers

In this section, we investigate how users interact with answers of different quality. Do users spend more time reading higher quality answers? Do they view more answers if they are of higher quality? How deeply do they explore the content, and does the depth of exploration depend on the quality? It is reasonable to doubt that editorial quality assessments agree with those of users, or are indicative of how interesting the content to users is. The behavioral data reported in this section directly answers these questions.

Experimental Setting and Instrumentation

In order to investigate how users view the content, a small fraction of page views on desktop in Yahoo Answers were instrumented. In this experiment, we focused on users who arrive to the CQA content by referral from a search engine's result page (typically, these are not the original asker or answerers). For each page view in the sample, we tracked *dwell time* (time on page from entry to exit) and *scrolling* (needed to expose more content). In the desktop interface, the answers to a question are arranged vertically; the question and the best answer (if any) are shown at the top, as well as between 2 and 4 other answers, depending on answer length and screen resolution. The answers are arranged in order of decreasing AQS. To expose additional answers, if they exist, scrolling is required. For this experiment, we recorded each scroll event as well as the maximum scroll depth (maximum pixel position of scroll marker). We used the *maximum scroll position* as a rough proxy for the content the user was willing to explore, and *dwell time* as a proxy for the users' interest. To make the analysis more meaningful, we split the page views into those with "high" and "low" AQS of the top-ranked answer, as that is the one guaranteed to be available and likely to be examined by the users. The "high" threshold for AQS was chosen as the median AQS for all answers in the dataset, and the "low" threshold was set to the 25% lowest quantile of the answers in the data. Other thresholds were experimented with for sensitivity analysis, without noticeable change to the reported statistics.

Overall Engagement Statistics

The overall dataset and engagement statistics are summarized in Table 1. In the dataset, there was a considerably larger number of pages with high-quality scores for the top answer than with low-quality scores. We conjecture that this is because we only considered search-intent page views, which privileged high-quality question-and-answer documents. Nevertheless, there was a substantial amount of page views for pages with low AQS of the top answer. The average dwell time on pages with high AQS (261 seconds) was more than a minute longer than for low AQS pages (158 seconds), suggesting that users are paying more attention to the higher quality content. Interestingly, the fraction of the time the users scroll to expose additional answers decreased for both high AQS (58% of page views) and low AQS (32% of page views). While users are almost twice as likely to explore additional answers when the top answer is of high quality, at first glance, the reduction in scrolling compared to the rest of the page views is puzzling. We conjecture, however, that we are observing two different phenomena. In the case of high-quality AQS pages, reduction in scroll is likely due to searcher satisfaction [29]: the searcher is more likely to be *satisfied* with the best, high-quality answer – not requiring her to explore additional answers (hence, no scrolling is needed). In the case of low-quality AQS pages, the dramatic reduction in scrolling behavior is likely due to the different phenomena of the searcher *abandoning* the page, as the examined top answer is of poor quality, and the searcher does not expect to find additional good content lower down. These overall behavioral results obtained agree with the A/B testing and manual annotation findings described in the previous sections.

To gain more precise understanding of how content quality affects exploration and engagement, we set out to inspect the scrolling behavior in more detail.

Scrolling Behavior and Exploration Depth

We now investigate in more detail whether answers of higher quality lead users to explore the content in more depth, and how this changes with the number of available answers. Figure 4 shows the average maximum scroll depth for varying number of *high-quality* answers (on the x-axis). In order to control for the total number of available answers, we separately report the scroll depth for different subsets of the page views, with 3, 4, 5, 6, and 7 answers available of any quality.

[5]A thoroughly complete test would also include a version of the treatment with just the low-quality answers hidden; but the engineering cost of this was prohibitive.

Statistic	All	Questions with high AQS score	Questions with low AQS score
Page views	108,787	49,721 (45.7%)	4,702 (4.3%)
Dwell time (seconds)	232	261 (+12.5%)	158 (-32%)
Views with scrolling (%)	75.5%	58.0% (-23.2%)	31.8% (-57.8%)

Table 1. User engagement statistics for a sample of questions with High and Low AQS scores for their top answer.

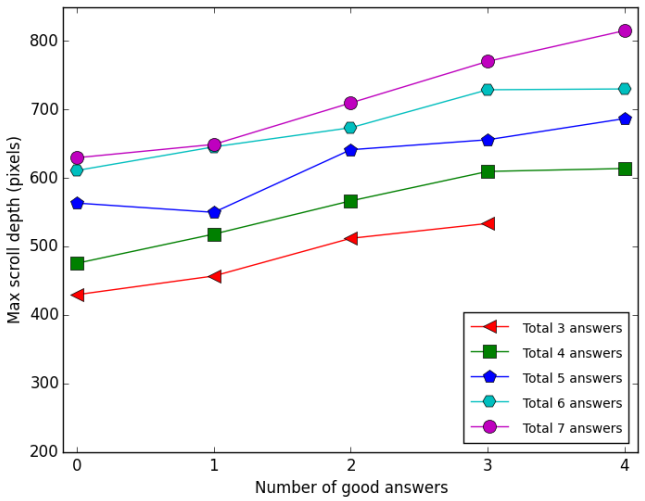

Figure 4. Average scroll depth for page views with varying number of high-quality answers, separated by total number of answers available.

It can be seen that the scroll depth is higher for pages with a larger number of total answers available, which serves as a "sanity check" of the data. More important is the difference within each data series. Consistently, the more answers with high-quality score presented, the more deeply users explored the page. For example, for pages with a total of 4 answers (green square markers), users scrolled, on average, 475 pixels down when all the answers were of low quality, compared to 615 pixels on average when all 4 answers were of high quality. This indicates that the factor at play here is content quality, rather than quantity.

DISCUSSION AND FUTURE WORK

We studied the deployment of an algorithmic quality scoring system within a large CQA site. We made an effort to build the best AQS possible given the rich prior art on the topic, taking advantage of many different features for high-quality content that have already been identified in various papers [1, 4, 7, 19, 21, 37]. We also added four novel features, one of which (ESA-based question-answer similarity [11]) was found to be among the most important features. We did not compare the performance of our AQS to those of previous studies, as this is not the focus of this work — we do not aim to improve the state-of-the-art in AQS, only to make use of it for our research.

Table 2 summarizes the five main experiments we conducted and their key results. The first two experiments focused on quality evaluation. We compared the best answer as determined by the AQS algorithm to two principal baselines: the best answer as selected by the asker and the best answer by the crowd's votes. While there was a high agreement between

Type	Evaluation Description	Key Result(s)
Quality	Comparing AQS to best answer by asker	63% agreement; AQS better in 68% of non-tie cases
Quality	Comparing AQS to best answer by "crowd" (votes)	29% agreement; AQS better for questions with 20 votes or less (57% of non-tie cases)
Engagement	Measuring clicks on live mobile traffic, with and without AQS treatment	AQS treatment improves CTR (9.2%) and MRR (3.8%)
Engagement	Dwell time for high-AQS vs. low-AQS answer results in desktop answer search	Average dwell time is substantially longer for high-AQS results
Engagement	Scroll depth for high-AQS vs. low-AQS answer results in desktop answer search	Users scroll deeper when high-AQS answers are presented

Table 2. Summary of key experimental results.

the AQS best answer and the asker's choice (when available), the AQS best answer was indicated to be better in the majority of the non-tie cases. The agreement between the AQS best answer and the best answer by votes was much lower, while in the non-tie cases, the AQS best answer was indicated to be better for questions with 20 votes or less, which account for 99% of the entire dataset. These results illustrate that a "thumb"-type reaction may only reflect a short-lived sentiment of the user giving the feedback (be it the asker, or a viewer) on the content, but it has limited power to predict or improve how future viewers judge it. Therefore, algorithmic scoring is a better tool to rank content for viewing which, in itself, is one significant result of this work.

One could think, however, that the 1% of questions with more than 20 votes are so popular, that they get the majority of the page views. To further explore this, we inspected the overall distribution of Yahoo Answers page views. We found that the portion of views for questions with over 20 votes is slightly less than 2% of the total set of question page views. This indicates that while the number of views for questions with more than 20 votes is somewhat higher than for the rest of the questions, they still account for a very small portion of the entire set of question page views.

Our additional set of experiments, summarized at the bottom of Table 2, focused on the effects of AQS-based filtering on user engagement [13, 27]. These experiments showed that AQS treatment improves the answer click-through rate and click position and that high AQS answers are explored in more depth, reflected both in the dwell time dedicated to clicked answers and the scroll depth of the entire answer result page. These results validate that algorithmic quality scores do not just agree with the editorial assessments, but also predict maintained user interest and engagement. To our knowledge, this is the first large scale *behavior and engage-*

ment validation of automatic quality assessment of user generated content, showing that indeed high algorithmic scores attract user attention, result in higher engagement, and encourage users to persist in exploring content deeper.

Our experiments involved analyzing user behavior data in a live product. All the experiments, including the analyzed data, reported results, and, when relevant, online intervention, were approved by a small committee that included the product manager and made sure the experiments were in line with the product's terms of service. The only experiment that involved real-time intervention was the click position experiment, however these interventions (re-ranking and truncating of answer results) were not expected to impose significant stress over users. In any case, our experiments did not include the collection of any personal data and did not involve personalized features. All statistics were collected in an aggregate way, without involving any user specific data.

When approving the research, the committee also took into account available resources on the engineering side that could support it. In our case, the answer exploration experiments were approved for the desktop platform, while the click position experiment was approved for the mobile platform. We did not see this as harmful for our experimentation; on the contrary, it allowed us to experiment with both types of available user interfaces and increase the overall diversity of our research.

Our experiments were conducted within Yahoo Answers and the results are naturally influenced by the characteristics of this specific site. Yet, the features we used for our AQS algorithm, as well as the concepts we used for evaluation (asker's selected best answer; up-votes and down-votes; clicked answers; dwell time and scroll depth), are not specific to Yahoo Answers and are available in additional major CQA sites, such as Stack Overflow [7] and Quora [40]. We therefore expect similar effects in other CQA sites, given the availability of a corresponding AQS system.

A number of directions for investigation remain. Our comparison between AQS and crowd was based on explicit crowd feedback in the form of public votes. Other types of implicit feedback, such as clicks, mouse movements, or scrolling, could be used to further refine crowd-based ranking [23]. Although sometimes used by popular websites, this type of implicit feedback is often hard to obtain and is also sensitive from a privacy perspective. Future work may examine whether algorithmic-based quality scoring poses similar value when compared against richer feedback from the crowd that includes implicit signals.

While there is, overall, strong evidence that AQS curation improves user engagement, more fine-grained analysis is possible. For instance, automatic scoring might be more aligned with user engagement for information-oriented questions compared to conversational questions [15], which could be worthwhile investigating in more detail. More generally, user-generated content is personal, and subjective criteria for content quality may naturally vary for different users, or even for different information needs. Thus, personalizing both AQS curation and presentation techniques could further improve user engagement and satisfaction, and remains to be explored in future work.

CONCLUSIONS

This work describes the deployment of an end-to-end automated content quality filtering system in a large community question-answering site with millions of users, and an even larger number of visitors arriving via search engines to find information.

The implementation of the quality filtering algorithm builds on prior work on content quality scoring, and progresses it to a deployment on a massive-scale web site with an existing corpus of hundreds of millions of items and a steady inflow of even more. Along the way, the system-generated scores are evaluated, both automatically and manually, to show that they align with a standard notion of quality. In the course of evaluation, we arrive at two new insights: (1) the algorithmically-chosen best answers are better than those chosen by the asker, and (2) they also outperform crowd sentiment, for crowds of up to 20 individuals.

Finally, and perhaps most importantly, this work closes the loop by showing improved user engagement on live traffic to the site. This is done both at the content selection level, showing user preference for curated content, and also at the scroll and dwell time levels, showing more satisfying interactions. This part is needed to prove that the preceding steps are more than an in-vitro exercise which has little connection with the needs of the actual users. To our knowledge, this paper is the first to investigate these effects at such large scale in the context of community question-answering.

ACKNOWLEDGMENTS

This work would not have been possible without the work of the Yahoo engineering team, including Somesh Jain, Archit Shrivastava, and Rajiv Verma; the product management of Shirin Oskooi; and the instrumentation work by Arun Kumar and Yash Dayal. We thank Rameez Akbar for overseeing the human evaluation work. We are indebted to David Carmel and Yuval Pinter for developing the AQS framework. We also wish to thank Mounia Lalmas for helpful insights in the early stages and to Mark Shovman for advising on statistics.

REFERENCES

1. Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding High-quality Content in Social Media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM '08)*. ACM, New York, NY, USA, 183–194.
2. Kohei Arai and Anik Nur Handayani. 2013. Predicting quality of answer in collaborative Q/A community. *Society and culture* 377993 (2013), 37799.
3. Jaime Arguello and Kyle Shaffer. 2015. Predicting Speech Acts in MOOC Forum Posts.
4. Jiang Bian, Yandong Liu, Ding Zhou, Eugene Agichtein, and Hongyuan Zha. 2009. Learning to Recognize

Reliable Users and Content in Social Media with Coupled Mutual Reinforcement. In *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*. ACM, New York, NY, USA, 51–60.

5. Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM, 675–684.

6. Steve Cronen-Townsend, Yun Zhou, and W Bruce Croft. 2002. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 299–306.

7. Daniel Hasan Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. 2013. Exploiting user feedback to learn to rank answers in q&a forums: a case study with stack overflow. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 543–552.

8. Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. 2009. How opinions are received by online communities: a case study on Amazon.com helpfulness votes. In *Proceedings of the 18th international conference on World wide web*. ACM, 141–150.

9. Georges Dupret and Mounia Lalmas. 2013. Absence time and user engagement: evaluating ranking functions. In *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 173–182.

10. Jill Freyne, Michal Jacovi, Ido Guy, and Werner Geyer. 2009. Increasing Engagement Through Early Recommender Intervention. In *Proceedings of the Third ACM Conference on Recommender Systems (RecSys '09)*. ACM, New York, NY, USA, 85–92.

11. Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis.. In *IJCAI*, Vol. 7. 1606–1611.

12. Anindya Ghose and Panagiotis G Ipeirotis. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *Knowledge and Data Engineering, IEEE Transactions on* 23, 10 (2011), 1498–1512.

13. Qi Guo and Eugene Agichtein. 2012. Beyond Dwell Time: Estimating Document Relevance from Cursor Movements and Other Post-click Searcher Behavior. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*. ACM, New York, NY, USA, 569–578.

14. Qi Guo, Haojian Jin, Dmitry Lagun, Shuai Yuan, and Eugene Agichtein. 2013. Mining touch interaction data on mobile devices to predict web search result relevance. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 153–162.

15. F. Maxwell Harper, Daniel Moy, and Joseph A. Konstan. 2009. Facts or Friends?: Distinguishing Informational and Conversational Questions in Social Q&A Sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 759–768.

16. F Maxwell Harper, Daphne Raban, Sheizaf Rafaeli, and Joseph A Konstan. 2008. Predictors of answer quality in online Q&A sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 865–874.

17. Ahmed Hassan and Ryen W White. 2013. Personalized models of search satisfaction. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2009–2018.

18. Wassily Hoeffding. 1963. Probability Inequalities for Sums of Bounded Random Variables. *J. Amer. Statist. Assoc.* 58, 301 (March 1963), 13–30.

19. Haifeng Hu, Bingquan Liu, Baoxun Wang, Ming Liu, and Xiaolong Wang. 2013. Multimodal DBN for Predicting High-Quality Answers in cQA portals. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, 843–847.

20. Nan Hu, Jie Zhang, and Paul A Pavlou. 2009. Overcoming the J-shaped distribution of product reviews. *Commun. ACM* 52, 10 (2009), 144–147.

21. Jiwoon Jeon, W. Bruce Croft, Joon Ho Lee, and Soyeon Park. 2006. A Framework to Predict the Quality of Answers with Non-textual Features. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. ACM, New York, NY, USA, 228–235.

22. Tapas Kanungo and David Orr. 2009. Predicting the readability of short web summaries. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. ACM, 202–211.

23. Diane Kelly and Jaime Teevan. 2003. Implicit Feedback for Inferring User Preference: A Bibliography. *SIGIR Forum* 37, 2 (Sept. 2003), 18–28.

24. J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Technical Report. DTIC Document.

25. Jon M Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* 46, 5 (1999), 604–632.

26. Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. 2014. Towards Better Measurement of Attention and Satisfaction in Mobile Search. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. ACM, New York, NY, USA, 113–122.
27. Janette Lehmann, Mounia Lalmas, Elad Yom-Tov, and Georges Dupret. 2012. Models of user engagement. In *User Modeling, Adaptation, and Personalization*. Springer, 164–175.
28. Chao Liu, Ryen W White, and Susan Dumais. 2010. Understanding web browsing behaviors through weibull analysis of dwell time. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 379–386.
29. Qiaoling Liu, Eugene Agichtein, Gideon Dror, Evgeniy Gabrilovich, Yoelle Maarek, Dan Pelleg, and Idan Szpektor. 2011. Predicting web searcher satisfaction with existing community-based answers. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 415–424.
30. Yandong Liu, Jiang Bian, and Eugene Agichtein. 2008. Predicting information seeker satisfaction in community question answering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR)*. 483–490.
31. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.
32. Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146 (2013), 647–651.
33. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Stanford InfoLab. Previous number = SIDL-WP-1999-0120.
34. Tetsuya Sakai, Daisuke Ishikawa, Noriko Kando, Yohei Seki, Kazuko Kuriyama, and Chin-Yew Lin. 2011. Using Graded-relevance Metrics for Evaluating Community QA Answer Selection. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11)*. ACM, New York, NY, USA, 187–196.
35. Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311, 5762 (2006), 854–856.
36. Chirag Shah, Sanghee Oh, and Jung Sun Oh. 2009. Research agenda for social Q&A. *Library & Information Science Research* 31, 4 (2009), 205–209.
37. Chirag Shah and Jefferey Pomerantz. 2010. Evaluating and Predicting Answer Quality in Community QA. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)*. ACM, New York, NY, USA, 411–418.
38. Maggy Anastasia Suryanto, Ee Peng Lim, Aixin Sun, and Roger H. L. Chiang. 2009. Quality-aware Collaborative Question Answering: Methods and Evaluation. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM '09)*. ACM, New York, NY, USA, 142–151.
39. Chenhao Tan, Evgeniy Gabrilovich, and Bo Pang. 2012. To each his own: personalized content selection based on text comprehensibility. In *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 233–242.
40. Gang Wang, Konark Gill, Manish Mohanlal, Haitao Zheng, and Ben Y. Zhao. 2013. Wisdom in the Social Crowd: An Analysis of Quora. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1341–1352.
41. Xin-Jing Wang, Xudong Tu, Dan Feng, and Lei Zhang. 2009. Ranking Community Answers by Modeling Question-answer Relationships via Analogical Reasoning. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*. ACM, New York, NY, USA, 179–186.
42. Guangyou Zhou, Kang Liu, and Jun Zhao. 2012. Joint Relevance and Answer Quality Learning for Question Routing in Community QA. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 1492–1496.
43. Yun Zhou and W Bruce Croft. 2007. Query performance prediction in web search environments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 543–550.