Analyzing and Predicting Question Quality in Community Question Answering Services

Baichuan Li[1], Tan Jin[1], Michael R. Lyu[1], Irwin King[2,1], and Barley Mak[1]
[1]The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
[2]AT&T Labs Research, San Francisco, CA, USA
bcli@cse.cuhk.edu.hk, tjin@cuhk.edu.hk, lyu@cse.cuhk.edu.hk,
irwin@research.att.com, king@cse.cuhk.edu.hk, barleymak@cuhk.edu.hk

ABSTRACT

Users tend to ask and answer questions in community question answering (CQA) services to seek information and share knowledge. A corollary is that myriad of questions and answers appear in CQA service. Accordingly, volumes of studies have been taken to explore the answer quality so as to provide a preliminary screening for better answers. However, to our knowledge, less attention has so far been paid to question quality in CQA. Knowing question quality provides us with finding and recommending good questions together with identifying bad ones which hinder the CQA service. In this paper, we are conducting two studies to investigate the question quality issue. The first study analyzes the factors of question quality and finds that the interaction between askers and topics results in the differences of question quality. Based on this finding, in the second study we propose a Mutual Reinforcement-based Label Propagation (MRLP) algorithm to predict question quality. We experiment with Yahoo! Answers data and the results demonstrate the effectiveness of our algorithm in distinguishing high-quality questions from low-quality ones.

Categories and Subject Descriptors

H.3.4 [**System and Software**]: question answering (fact retrieval) systems; H.3.5 [**Online Information Services**]: Web-based services

General Terms

Algorithms, Measurement, Experimentation

Keywords

Community Question Answering, Question Quality, Analysis, Prediction

1. INTRODUCTION

Community Question Answering (CQA) services provide a platform for users to ask and answer questions covering a wide range of topics. Different with traditional Question Answering (QA) using stored data to answer questions automatically using natural language, users in CQA ask and answer questions to seek information and share knowledge by themselves. Recently, an increasing number of users are choosing CQA to solve problems, for example, Both Yahoo! Answers[1] and Baidu Zhidao[2] have more than 10 million daily visits in 2011 according to Google Trends[3]. Popularity of CQA, however, brings about a huge amount of questions and answers. Series of studies are taken to investigate the answer quality in CQA so as to screen for better answers [12, 1, 21, 20, 16], but as for question quality, fewer studies have so far been documented. In fact, questions in CQA vary in attracting user attention, answering attempts and the best answer. Taking questions in Yahoo! Answers as an example, some questions acquire thousands of tag-of-interests and answering attempts while some questions fail to get any answering attempts, indicating varied degrees of question quality.

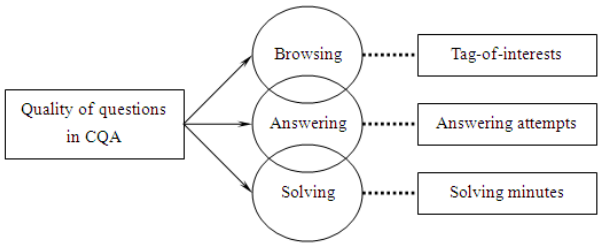


Figure 1: Construct of question quality in CQA

The significance of finding question quality in CQA lies in these four points: (1) **Question quality affect answer quality**. It is observed that low quality questions always lead to bad answers while high quality questions usually receive good answers [1]. (2) **Low quality questions hinder the CQA service**. Low quality questions, such as commercial advertisements, reduce the user experience greatly. (3) **High quality questions promote the development of the community**. Since high quality questions attract more users to contribute their knowledge, they not only improve the efficiency of solving questions but also enrich the knowledge base of the community. (4) **Question quality facilitates question finding**. We will improve question retrieval and question recommendation in CQA services if taking question quality into account.

Text quality, such as "accuracy and comprehensiveness" (see [4]) has been widely applied to assess answer quality

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

[1]http://answers.yahoo.com/
[2]http://zhidao.baidu.com/
[3]http://trends.google.com/

but is not appropriate to estimate question quality because well-written tangible texts contribute little about question quality. In this paper, we use the term "question quality" to represent the question's "social quality", which involves three dimensions (see Fig. 1): (1) user attention; (2) answering attempts; and (3) best answer. In other words, "high quality" questions are supposed to attract great user attention, more answer attempts and receive best answers within a short period. Otherwise, questions failing to achieve the three criteria are labeled as "low quality" questions since the questions neither meet user needs nor contribute to the knowledge base of the community.

The paper has six sections. We first review related work in Section 2. Then, in Section 3, we present the experimental data and the ground truth. Next, two studies are reported in Sections 4 and 5 respectively. Study one applies statistical analyses to find factors affecting the question quality. Based on the findings of study one, study two proposes a novel graph-based Semi-supervised Learning (SSL) algorithm and applies the algorithm to predict the question quality. We conclude the paper in Section 6.

2. RELATED WORK

Content quality prediction and evaluation in CQA service and label propagation on graphs are two research topics pertinent to our work.

Content quality prediction and evaluation. The current studies of questions in CQA service focus on question retrieval [5, 6, 22, 3, 9] and question recommendation [24, 7, 17, 13, 14]. However, less work deals with evaluating or predicting question quality. Agichtein et al. [1] first analyze essential features to the quality of questions, where question quality is defined as "well-formedness, readability, utility, and interestingness". Afterwards, Bian et al. [4] link the relationship among users, questions and answers to estimate question quality, answer quality and user expertise. These two studies have laid conceptual and methodological foundations for this paper. In this paper we define question quality from perspectives of users and community development, also taking account of contribution of questions.

Answer quality in CQA, on the other hand, has been widely investigated in the past few years and researchers have been working to distinguish good answers from bad ones, facilitating users with asking questions in CQA. One of the typical ways is ranking answers using answer features. Jeon et al. [12] specify non-textual features for CQA answer quality prediction and Agichtein et al. [1] leverage more features like community feedback to identify high quality content. Recently, Sakai et al. [18] propose to employ graded-relevance metrics to evaluate answer quality.

At the same time, ranking algorithms and models are explored as well. Bian et al. [3] rank answers of factual information retrieval according to user interaction, answer quality and relevance. Wang et al. [21] devise an answer ranking algorithm which applies analogical reasoning to model the relation between questions and answers in CQA. Suryanto et al. [20] construct models to find good answers of new questions from a CQA portal considering user expertise in answering. The coupled mutual reinforcement model proposed by Bian et al. [4] is most related to our method. In their model, question quality is determined by answer quality and asker expertise, which echo our claim of question quality construct covering user attention, answering attempts and best answer. However, in their framework question quality is estimated directly from answer quality and asker expertise, but our method is to predict question quality without any answer information.

Label propagation. Our proposed algorithm is an extension of label propagation on bipartite graphs. Label propagation is a class of algorithms which propagates the labels of labeled data to unlabeled data in a homogeneous graph. Harmonic function [26], local and global consistency [23] and green's function [8] are three typical label propagation methods. Our algorithm is based on the harmonic function [26], which assumes the label of each unlabeled data is the weighted average of its neighbors'.

Table 1: Summary of questions and askers in *Entertainment & Music* category and its subcategories

Subcategory	# of questions	# of askers
Celebrities	11,817	7,087
Comics & Animation	11,327	6,801
Horoscopes	7,235	2,203
Jokes & Riddles	3,685	2,569
Magazines	548	462
Movies	15,121	10,996
Music	32,948	18,589
Other - Entertainment	2,244	2,003
Polls & Surveys	138,507	18,685
Radio	640	272
Television	14,477	10,146
All	238,549	62,853

3. DATA DESCRIPTION

In this section, we first describe our data set. Then we detail how to set the ground truth for question quality, providing the baseline for the following studies and analyses.

3.1 Data set

We collect 238,549 resolved questions from July 7, 2010 to September 6, 2010 under the *Entertainment & Music* category of Yahoo! Answers. For each question, we crawl both the question information (the texts of subject and content, post time, best answer post time, number of answers and number of tag-of-interests by other users) and the asker information (total points, # of answers, # of best answers, # of questions asked, # of questions resolved and # of stars received). There are altogether 11 subcategories under *Entertainment & Music* and Table 1 gives the statistics of the data set.

3.2 Ground truth

We set the ground truth using the construct of question quality in CQA (see Fig. 1). To quantify the three variables, we are using the number of tag-of-interests (NT, reflecting the attractiveness of a question), number of answers (NA), and the reciprocal of the minutes for getting the best answer (RM) in this paper.

We first attempt to cluster these questions but the cluster-

Table 2: Rule base for the ground truth setting

RM \ NTA	4	3	2	1
4	4	4	3	2
3	4	3	3	2
2	3	3	2	1
1	2	2	1	1

Table 3: Summary of questions in four levels

Level	1	2	3	4
Count	53,806	62,192	69,836	52,715

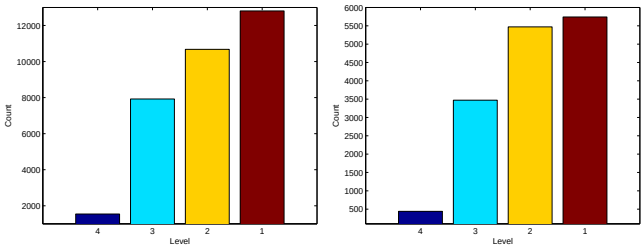
ing results are not congruent with different seeds. In spite of this, the size of each cluster varies sharply from less than 10 to more than 50,000. Having consulted domain experts, we resort to expert based reasoning. The Pearson Correlations between each of the two variables are calculated and NT and NA are correlated (0.500) but either NT or NA shows little correlation with RM (-0.011 and 0.213 respectively). Therefore, we first normalize and average the values of NT and NA and then convert them into an integer in a scale from 1 to 4 (NTA hereafter, with 4 the highest quality) using three equidistant cutting points of 0.75 (top 25%), 0.50 and 0.25 to assign each band roughly the same amount of questions. At the same time, RM is also transformed into 1 to 4 scale data using such approach. After that, two scale data are reasoned based on the rule base (see Table 2), which comes from consensus among the authors and domain experts. In the end, all questions are labeled as from level 1 to level 4, with level 4 the highest quality questions. Table 3 summarizes questions with levels and they are taken as the ground truth.

4. STUDY ONE: FACTORS AFFECTING QUESTION QUALITY

In CQA portals, askers are posting questions on different topics and as such askers and topics are probably the main sources of varied question quality. However, we know little about the contribution of askers and topics to the question quality. Here, we are concerning which factors have the major impacts on question quality and we use the subcategories under *Entertainment & Music* as various topics. We do not select different categories as topics in that: first, we observe that the majority of users only ask questions in a very few categories, thus choosing subcategories as topics are more representative; second, different subcategories also reflect various topics, for instance, music, movies, polls and surveys are three distinctive aspects of entertainment.

Study one is designed thus. We first select the two most popular subcategories[4] (namely, *Music* and *Movies*, see Table 1) as two representative topics in study one and then check their distributions of question quality. Next, we track

[4]The subcategory *Polls & Surveys* is not chosen since this subcategory is used to elicit public opinion and we observe questions in this subcategory usually receive much more answers than others.



(a) (b)

Figure 2: Distributions of question quality in three topics. (a) Music; (b) Movies.

Table 4: Summary of question quality for different askers.

User	Music		Movies	
	Mean	Std	Mean	Std
1	2.50	0.93	2.17	0.41
2	2.45	0.52	2.57	0.98
3	1.86	0.90	1.45	0.82
4	2.65	0.72	2.60	0.55
5	1.90	0.74	2.00	0.71
6	2.62	0.87	1.83	0.86
7	2.48	0.68	2.20	0.84
8	2.86	0.92	2.14	0.90
9	2.38	0.92	2.30	1.06
10	2.50	0.53	2.40	0.55
11	2.00	0.71	1.50	0.55
12	2.48	0.95	2.47	0.84
13	2.84	0.68	2.83	0.41
14	1.33	0.52	2.40	0.89
15	1.90	0.74	1.83	0.75
16	1.80	0.84	1.83	0.75
17	2.15	0.55	2.50	1.05
18	2.36	0.92	1.67	0.87
19	2.00	1.00	2.00	1.00
20	2.00	0.67	2.00	1.00
21	2.69	0.68	2.80	0.45
22	2.13	0.99	2.57	1.27

askers with at least five questions in both these two subcategories and test question quality of these questions.

Figure 2 presents the histograms of question quality of *Music*, and *Movies*. We can find that the distributions of question quality in *Music* and *Movies* are close: the number of questions increases with question quality decreases from level four to level one; the proportions of each level's questions are similar. The difference lies in that the proportion of questions in level two of *Movies* is larger. This observation tells us topics only cannot distinguish good questions from bad ones.

To investigate the influence of askers, we select a total of 22 askers who have asked at least 5 questions in the two sub-categories. Mean and standard deviation of the question quality are reported in Table 4. Our observations are: 1) Different askers own various question qualities at the same topic. For instance, question quality of user 8 is much higher than that of user 16; 2) The question quality of the same asker on various topics have great differences. E.g., user 14 asks many good questions about *Movies*, but his/her question quality in *Music* is poor. Therefore, we find that it is the interaction between asker and topics which plays the

most import role in distinguishing good questions from bad ones.

To sum up, study one examines the effects of askers and topics on question quality. We observe that topics themselves cannot determine question quality, and the interaction between askers and topics is the most important factor affecting question quality. This observation motivates us to design a novel algorithm to predict question quality in the next study.

5. STUDY TWO: PREDICTION OF QUESTION QUALITY

Study one has uncovered the main factors of question quality, but it is taken place when questions are resolved. In study two, we have an even more challenging prediction work: estimating question quality right after a question is posted but still not answered by any answerers. Motivating by the result of study one, we model the relationships among questions, topics and askers as a bipartite graph model. Figure 3 shows one example, where u_1, u_2, and u_3 asks five questions $(q_1, ..., q_5)$ in three topics $(t_1, t_2,$ and $t_3)$. Each edge linking an asker and a question represents the question asked by the asker and each rectangle denotes a topic. In the example, we know that u_1 asks q_1 and q_3, and q_2 is in topic t_1. Here topics are represented by subcategories or categories in CQA portals.

The ideas of our algorithm are straightforward:

1. As for the same topics, questions with similar structures and expressions will have identical quality and users with same profiles will embrace approximate asking expertise.

2. As for different topics, users' abilities to ask good questions are not equivalent and such abilities are constant within a particular period.

3. Each question's quality is estimated from the qualities of similar questions and the asker's abilities to ask good questions in that topic. Meanwhile, each asker's ability of asking good questions at one topic is estimated from his/her question quality and similar askers' asking abilities in that topic.

Based on the these, we propose a graph-based SSL algorithm called "Mutual Reinforcement Label Propagation" (**MRLP**) to predict question quality in CQA service. Before introducing **MRLP**, we first give the formal definitions of question quality and users' asking expertise.

Definition 5.1 (**Question quality**). *Question q_i's quality is represented by $\hat{q}_i$, which refers to its ability to attract user attention, get answering attempts and receive the best answer efficiently. It ranges from 0 to 1. The higher value is, the higher quality the question has.*

Definition 5.2 (**Asking expertise**). *User u_j's asking expertise in topic t_k is represented by $\hat{u}_{jk}$, which reflects the user's ability to ask high quality questions within that topic. $\hat{u}_{jk}$ ranges from 0 to 1. It is worth noting that $\hat{u}_{jk}$ models the effect of interaction of the asker and the topic.*

5.1 MRLP

Suppose there are m askers who ask n questions in t topics, let $U^1, U^2, ..., U^t$ denote the vectors $(m \times 1)$ of askers' asking

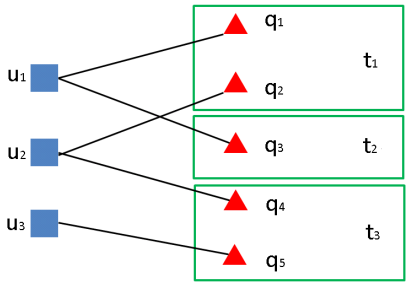


Figure 3: A toy example. Left: askers; Right: questions in various topics.

expertise in these topics, and $Q(n \times 1)$ denote the vector of question quality, we define a $m \times n$ matrix E, where $e_{ij} = 1(i \in [1, m], j \in [1, n])$ means u_i asks q_j, otherwise $e_{ij} = 0$. From E we get E':

$$E'_{ij} = \frac{e_{ij}}{\sum_{k=1}^{n} e_{ik}}. \quad (1)$$

For the question part of the bipartite graph, we create edges between any two questions within same topics. The weight for the edge linking q_i and q_j is represented by $w(q_i, q_j)$, which is calculated from the cosine similarity between the features of two questions $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$w(q_i, q_j) = \exp(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\lambda_q^2}), \quad (2)$$

where λ_q is a weighting parameter. $w(q_i, q_j)$ is set to be 0 if q_i and q_j belong to two different topics. In addition, we define $w(q_i, q_i) = 0$.

Then, we define an $n \times n$ probabilistic transition matrix N:

$$N_{ij} = P(q_i \to q_j) = \frac{w(q_i, q_j)}{\sum_{k=1}^{n} w(q_i, q_k)}, \quad (3)$$

where N_{ij} is the probability of transit from q_i to q_j. Similarly, we create edges between any two askers who have asked questions in the same topic(s) for asker part of the graph with λ_a as the weighting parameter using Eq. (2). In addition, we define a $m \times m$ probabilistic transition matrix M like N in Eq. (3).

For topic t_k, given some known labels of U_k and/or Q, we describe the MRLP in Alg. 1. The equation at line 3 estimates users' asking expertise from their neighbors and their questions' qualities. Correspondingly, the equation at line 4 calculates questions' quality on topic k from their neighbors and their askers' asking expertise. Repeating **MRLP** k times, all questions' qualities and askers' asking expertise are estimated.

Now, we prove the convergence of the **MRLP**. Suppose there are l labeled data and u unlabeled data for questions' qualities together with x labeled data and y unlabeled data for askers' asking expertise, i.e., $Q^k = [\hat{q}_1^k, ..., \hat{q}_l^k, \hat{q}_{l+1}^k, ..., \hat{q}_{l+u}^k]^T$ and $U^k = [u_{1k}, ..., u_{xk}, u_{(x+1)k}, ..., u_{(x+y)k}]^T$. Thus, We can split E', E^T, M and N into four parts:

$$E' = \begin{bmatrix} E'_{xl} & E'_{xu} \\ E'_{yl} & E'_{yu} \end{bmatrix}, E^T = \begin{bmatrix} E^T_{xl} & E^T_{yl} \\ E^T_{xu} & E^T_{yu} \end{bmatrix},$$

$$M = \begin{bmatrix} M_{xx} & M_{xy} \\ M_{yx} & M_{yy} \end{bmatrix}, N = \begin{bmatrix} N_{ll} & N_{lu} \\ N_{ul} & N_{uu} \end{bmatrix}.$$

Thus, we get

$$\begin{bmatrix} U_x^k \\ U_y^k \end{bmatrix}_{c+1} = \alpha \begin{bmatrix} M_{xx} & M_{xy} \\ M_{yx} & M_{yy} \end{bmatrix} \begin{bmatrix} U_x^k \\ U_y^k \end{bmatrix}_c + (1-\alpha) \begin{bmatrix} E'_{xl} & E'_{xu} \\ E'_{yl} & E'_{yu} \end{bmatrix} \begin{bmatrix} Q_l^k \\ Q_u^k \end{bmatrix}_c,$$

and

$$\begin{bmatrix} Q_l^k \\ Q_u^k \end{bmatrix}_{c+1} = \beta \begin{bmatrix} N_{ll} & N_{lu} \\ N_{ul} & N_{uu} \end{bmatrix} \begin{bmatrix} Q_l^k \\ Q_u^k \end{bmatrix}_c + (1-\beta) \begin{bmatrix} E_{xl}^T & E_{yl}^T \\ E_{xu}^T & E_{yu}^T \end{bmatrix} \begin{bmatrix} U_x^k \\ U_y^k \end{bmatrix}_c.$$

Since U_x^k and Q_l^k are clamped to manual labels in each iteration, we now only consider U_y^k and Q_u^k. From the above two equations we get:

$$\begin{aligned} \begin{bmatrix} U_y^k \\ Q_u^k \end{bmatrix}_{c+1} &= \begin{bmatrix} \alpha M_{yy} & (1-\alpha) E'_{yu} \\ (1-\beta) E_{yu}^T & \beta N_{uu} \end{bmatrix} \begin{bmatrix} U_y^k \\ Q_u^k \end{bmatrix}_c \\ &+ \begin{bmatrix} \alpha M_{yx} U_x^k + (1-\alpha) E'_{yl} Q_l^k \\ \beta N_{ul} Q_l^k + (1-\beta) E_{xu}^T U_x^k \end{bmatrix}. \end{aligned}$$

Let

$$A = \begin{bmatrix} \alpha M_{yy} & (1-\alpha) E'_{yu} \\ (1-\beta) E_{yu}^T & \beta N_{uu} \end{bmatrix}, b = \begin{bmatrix} \alpha M_{yx} U_x^k + (1-\alpha) E'_{yl} Q_l^k \\ \beta N_{ul} Q_l^k + (1-\beta) E_{xu}^T U_x^k \end{bmatrix},$$

we get

$$\begin{bmatrix} U_y^k \\ Q_u^k \end{bmatrix}_n = A^n \begin{bmatrix} U_y^k \\ Q_u^k \end{bmatrix}_0 + (\sum_{i=1}^{n} A^{i-1}) b,$$

where $\begin{bmatrix} U_y^k \\ Q_u^k \end{bmatrix}_0$ are the initial values for unlabeled askers and questions. The following proof is similar to the one in Chapter 2 of [25]. Since M, N, E' and E^T are row normalized (each row of E^T only contains one "1", others are "0"), M_{yy}, N_{uu}, E'_{yu}, and E_{yu}^T are sub-matrixes of them,

$$\exists \gamma < 1, \sum_{j=1}^{y+u} A_{ij} \leq \gamma, \forall i = 1, ..., y+u.$$

So

$$\begin{aligned} \sum_j A_{ij}^n &= \sum_j \sum_k A_{ik}^{n-1} A_{kj} \\ &= \sum_k A_{ik}^{n-1} \sum_j A_{kj} \\ &\leq \sum_k A_{ik}^{n-1} \gamma \\ &\leq \gamma^n \end{aligned}$$

Therefore the sum of each row of A converges to zero, thus $A^n \begin{bmatrix} U_y^k \\ Q_u^k \end{bmatrix}_0 \to 0$. Finally we get

$$\begin{bmatrix} U_y^k \\ Q_u^k \end{bmatrix} = (I - A)^{-1} b,$$

which are fixed values.

5.2 Experimental setup

To verify the effectiveness of the **MRLP** in predicting question quality, we experiment with the data described in Section 3. For each topic of *Music* and *Movies*, we choose questions of those askers who asked at least 10 questions in that topic. Since our goal is to distinguish high quality questions from low quality ones, we follow the common binary classification setting in the previous work [19, 15, 1].

Algorithm 1 MRLP-ST

Input: user asking expertise vector U_0^k, question quality vector Q_0^k, E, transition matrixes M and N, weighting coefficients α and β, some manual labels of U_0^k and/or Q_0^k.

1: Set $c = 0$.
2: **while** not convergence **do**
3: Propagate user expertise. $U_{c+1}^k = \alpha \cdot M \cdot U_c^k + (1-\alpha) \cdot E' \cdot Q_c^k$.
4: Propagate question quality. $Q_{c+1}^k = \beta \cdot N \cdot Q_c^k + (1-\beta) \cdot E^T \cdot U_{c+1}^k$, where E^T is the transpose of E.
5: Clamp the labeled data of U_{c+1}^k and Q_{c+1}^k.
6: Set $c = c + 1$.
7: **end while**

Table 5: Summary of data in study two

	Music	Movies
# Questions	7,373	1,076
# High-Quality Questions	3,670	331
# Low-Quality Questions	3,703	745
# Askers	314	56

Thus, we take questions of level 3 and level 4 as high quality ones and the other questions as low quality ones. Table 5 summarizes the data. To get prediction performance at different training levels, we adjust the training rates from 10% to 90% in our experiments. For each rate we select the corresponding proportion of earlier posted questions as training data and the others as testing data.

5.2.1 Selected features

Referring to the work of [1] and [2], we adopt the features in Table 6 to construct graphs and train classifiers. They are divided into question-related and asker-related features. Question-related features are extracted from question text including subject and content; asker-related features come from askers' profiles. For features such as POS_entropy, we use the tool OpenNLP[5] to conduct tokenization, detect sentences and annotate the part-of-speech tags. In addition, we utilize the Microsoft Office Word Primary Interop Reference[6] to detect typo errors.

We also report the information gain of each feature in Table 6. It is found that all features' information gains are small, which means these features are not so salient to question quality. In addition, asker-related features are more crucial than question-related features since their information gains are higher. As for question-related features, space density and subject length are the most important ones.

5.2.2 Methods compared

We compare the MRLP with the following methods:

- **Logistic Regression**: Shah et al. [19] apply logistic regression model to predict answer quality in Yahoo! Answers. Here we adopt the same approach to predict question quality with question-related features only (**LR_Q**), and both question-related and asker-related features (**LR_QA**). These two methods are treated as baselines.

[5]http://opennlp.sourceforge.net/
[6]http://msdn.microsoft.com/library/bb406008(v=office.11).aspx

Table 6: Summary of features extracted from questions and askers

Name	Description	IG
Question-related features		
Sub_len	Number of words in question subject (title)	0.0115
Con_len	Number of words in question content	0.0029
Wh-type	Whether the question subject starts with Wh-word (e.g., "what", "where", etc.)	0.0001
Sub_punc_den	Number of question subject's punctuation over length	0.0072
Sub_typo_den	Number of question subject's typos over length	0.0021
Sub_space_den	Number of question subject's spaces over length	0.0138
Con_punc_den	Number of question content's punctuation over length	0.0096
Con_typo_den	Number of question content's typos over length	0.0006
Con_space_den	Number of question content's spaces over length	0.0113
Avg_word	Number of words per sentence in question's subject and content	0.0048
Cap_error	The fraction of sentences which are started with a small letter	0.0064
POS_entropy	The entropy of the part-of-speech tags of the question	0.0004
NF_ratio	The fraction of words that are not the top 10 frequent words in the collection	0.0009
Asker-related features		
Total_points	Total points the asker earns	0.0339
Total_answers	Number of answers the asker provided	0.0436
Best_answers	Number of best answers the asker provided	0.0331
Total_questions	Number of questions the asker provided	0.0339
Resolved_questions	Number of resolved questions asked by the asker	0.0357
Star_received	Number of stars received for all questions	0.0367

- **Stochastic Gradient Boosted Tree**: Agichtein et al. [1] report the stochastic gradient boosted trees [10] (SGBT) perform best among several classification algorithms including SVM and log-linear classifiers to classify content quality in CQA service. For SGBT classifier, in each iteration a new decision tree is built to fit a model to the residuals left by the classifier on the previous iteration. In addition, a stochastic element is added in each iteration to smooth the results and prevent overfitting. For different features we have **SGBT_Q** and **SGBT_QA**.

- **Harmonic Function**: Zhu et al. [26] propose the harmonic function algorithm for label propagation on a homogeneous graph, where all nodes (edges) represent the same kind of object (relationship). To estimate question quality, we create a graph in which each node stands for a question and each edge's weight represents two question's similarity. Let W denote the weight matrix and D denote the diagonal matrix with $d_i = \sum_j w_{ij}$, then construct stochastic matrix $P = D^{-1}W$. Let $f = \begin{bmatrix} f_l \\ f_u \end{bmatrix}$ where f_l are the qualities of labeled questions and f_u are what we want to predict. We split the matrix W (also D and P) into four parts:

$$W = \begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix},$$

where W_{ll} means the similarities among labeled questions, and W_{lu} means the similarities between labeled questions and unlabeled questions. The harmonic solution [26] is:

$$f_u = (D_{uu} - W_{uu})^{-1} W_{ul} f_l = (I - P_{uu})^{-1} P_{ul} f_l. \quad (4)$$

Similarly, we construct **HF_Q** and **HF_QA** using different features.

In our experiments we use the tool Weka [11] to build logistic regression models and SGBT classifiers. All parameters of these models are tuned through grid search using the data when training rate is 90%. Furthermore, we build 10-NN graphs for graph-based algorithms, i.e., HF_Q, HF_QA and MRLP.

5.2.3 Evaluation metrics

We adopt Accuracy, Sensitivity, and Specificity as the evaluation metrics. Accuracy reflects the overall performance of prediction, while Sensitivity and Specificity measure the algorithm's ability to classify high quality and low quality questions into correct classes respectively.

5.3 Experimental results

Table 7 reports the predicting accuracy of these methods under various training rates across three topics. Figures 4, 5, 6, and 7 present Sensitivity and Specificity of each method in *Music* and *Movies*.

From Table 7 we know the **MRLP** performs much better than baseline methods (**LR_Q** and **LR_QA**) in all settings. E.g., when training rate is 10% for *Movies*, the Accuracy of **MRLP** is 81.63% and 81.08% higher than that of **LR_Q** and **LR_QA**. In addition, **MRLP** is more effective in predicting question quality than other methods in most cases except when training rate is 10% for *Music* and 50% for *Movies*. This result demonstrates that **MRLP** are more effective in predicting questions' qualities through modeling the interaction between askers and topics and capturing the mutual reinforcement relationship between asking expertise and question quality.

Meanwhile, neither the **MRLP** nor other methods perform very well in classifying question quality across the two

Table 7: Different methods' performance with question-related features only versus both question-related and user-related features (*Music*: $\alpha = 0.2$, $\beta = 0.2$; *Movies*: $\alpha = 0.8$, $\beta = 0.1$)

Methods	Accuracy under training rate (%)									
	Music					Movies				
	10	30	50	70	90	10	30	50	70	90
LR_Q	0.542	0.442	0.440	0.428	0.415	0.381	0.332	**0.656**	0.384	0.546
LR_QA	0.552	0.439	0.442	0.442	0.408	0.376	0.333	0.652	0.387	0.519
HF_Q	0.535	0.528	0.545	0.541	0.547	0.541	0.487	0.548	0.514	0.546
HF_QA	0.535	0.545	0.559	0.557	0.565	0.505	0.496	0.548	0.499	0.491
SGBT_Q	0.550	0.574	0.590	0.592	0.576	0.595	0.570	0.548	0.539	0.593
SGBT_QA	**0.615**	0.605	0.631	0.634	0.637	0.595	0.527	0.587	0.542	0.593
MRLP	0.599	**0.612**	**0.633**	**0.656**	**0.664**	**0.607**	**0.603**	0.554	**0.582**	**0.611**

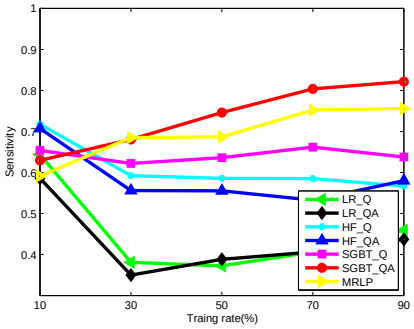


Figure 4: Sensitivity versus training rate across various methods in *Music*

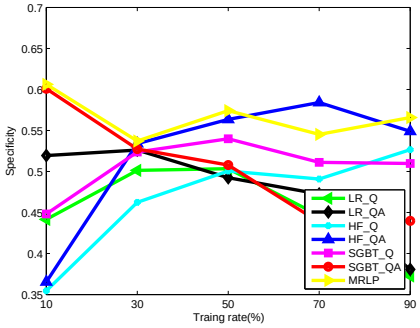


Figure 5: Specificity versus training rate across various methods in *Music*

topics. Even the training rate is set to be 90%, there are still more than 35% of questions not correctly classified. The reason is that question text and asker profile features are not salient features of question quality, as shown in Table 6. Since all features' information gains are less than 0.05, it is very hard to make satisfying prediction using these features.

5.3.1 *Question-related features vs. asker-related features*

Comparing **LR_Q**, **HF_Q**, and **SGBT_Q** with **LR_QA, HF_QA** and **SGBT_QA** from Table 7, we find that with asker-related features the accuracy of prediction is substantially higher than the same methods without using asker-related features in *Music*. However, there seems to be a decrease of accuracy if asker-related features are used in *Movies*, fewer askers in *Movies* may explain this special case. Figures 4, 5, 6, and 7 give more details. In specific, utilizing asker-related features increases the Sensitivity of **SGBT** and the Specificity of **LR** and **HF** in *Music*, and enhance the Sensitivity of **LR** and **HF** in *Movies*. However, it decreases the Sensitivity of **HF** and Specificity of **SGBT** in *Music* and the Specificity of **LR** and **HF** in *Movies*.

5.3.2 *Mixture vs. separation of user-related features*

Comparing **LR_QA, HF_QA** and **SGBT_QA** with **MRLP** which all use question-related and user-related features, **MRLP** performs the best on Accuracy. When looking at the Sensitivity in Fig. 4 and Fig. 6, the Specificity in Fig. 5 and Fig. 7, **MRLP** is more balanced in Sensitivity and Specificity than other algorithms. For instance, **LR_Q** has the highest Specificity for Movies but the lowest Sensitivity, which means it

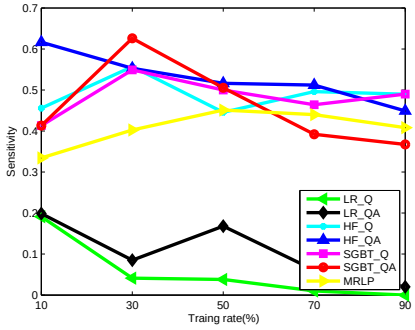


Figure 6: Sensitivity versus training rate across various methods in *Movies*

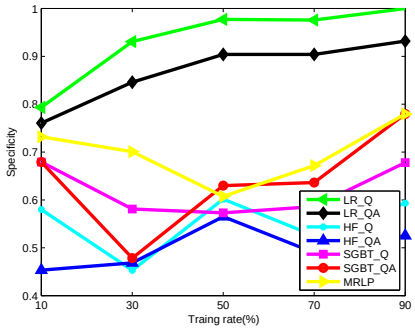


Figure 7: Specificity versus training rate across various methods in *Movies*

almost predicts all questions into low-quality ones. Thus, **MRLP** is more effective in discriminating high quality questions from low ones. Overall, **MRLP** gives the best performance since it integrates the question-related features with asker-related features naturally other than a simple combination. In particular, it improves the performance of the second best method (**SGBT_QA**) by 7% on average in *Music* and *Movies*. **MRLP** naturally separate question-related features and user-related features in graph construction, and the above results demonstrate this approach is better than simply combining these features.

6. CONCLUSION

In this paper, we conduct two studies to investigate question quality in CQA services. In study one, we analyze the factors influencing question quality and find that the interaction of users and topics leads to the difference of question quality. Based on the findings of study one, in study two we propose a mutual reinforcement-based label propagation algorithm to predict question quality using features of question text and asker profile.

We experiment with real world data set and the results demonstrate that our algorithm is more effective in distinguishing high quality questions from low quality ones than logistic regression model and other state-of-the-art algorithms, such as the stochastic gradient boosted tree and the harmonic function. However, as current features extracted from question text and asker profile are not so salient, neither our algorithm nor other classical methods achieves satisfactory performance at present.

Current results lead us to further explore the salient features of question quality in the future work. We also plan to utilize question quality to improve question search and question recommendation in CQA services.

7. ACKNOWLEDGEMENT

The work described in this paper was fully supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK 413210 and CUHK 415311) and two grants from Google Inc. (one for Focused Grant Project "Mobile 2014" and one for Google Research Awards).

8. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proc. of WSDM*, 2008.

[2] E. Agichtein, Y. Liu, and J. Bian. Modeling information-seeker satisfaction in community question answering. *ACM Trans. Knowl. Discov. Data*, 3(2):1–27, 2009.

[3] J. Bian, Y. Liu, E. Agichtein, and H. Zha. Finding the right facts in the crowd: factoid question answering over social media. In *Proc. of WWW*, 2008.

[4] J. Bian, Y. Liu, D. Zhou, E. Agichtein, and H. Zha. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *Proc. of WWW*, 2009.

[5] X. Cao, G. Cong, B. Cui, and C. S. Jensen. A generalized framework of exploring category information for question retrieval in community question answer archives. In *Proc. of WWW*, 2010.

[6] X. Cao, G. Cong, B. Cui, C. S. Jensen, and C. Zhang. The use of categorization information in language models for question retrieval. In *Proc. of CIKM*, 2009.

[7] S. D. Damon Horowitz. Anatomy of a large-scale social search engine. In *Proc. of WWW*, 2010.

[8] C. Ding, H. D. Simon, R. Jin, and T. Li. A learning framework using green's function and kernel regularization with application to recommender system. In *Proc. of KDD*, 2007.

[9] H. Duan, Y. Cao, C.-Y. Lin, and Y. Yu. Searching questions by identifying question topic and question focus. In *Proc. of ACL:HLT*, 2008.

[10] J. H. Friedman. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4):367–378, 2002.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009.

[12] J. Jeon, W. B. Croft, J. H. Lee, and S. Park. A framework to predict the quality of answers with non-textual features. In *Proc. of SIGIR*, 2006.

[13] B. Li and I. King. Routing questions to appropriate answerers in community question answering services. In *Proc. of CIKM*, 2010.

[14] B. Li, I. King, and M. R. Lyu. Question routing in community question answering: putting category in its place. In *Proc. of CIKM*, 2011.

[15] Y. Liu, J. Bian, and E. Agichtein. Predicting information seeker satisfaction in community question answering. In *Proc. of SIGIR*, 2008.

[16] J. Lou, K. Lim, Y. Fang, and Z. Peng. Drivers of knowledge contribution quality and quantity in online question and answering communities. In *Proc. of PACIS*, 2011.

[17] M. Qu, G. Qiu, X. He, C. Zhang, H. Wu, J. Bu, and C. Chen. Probabilistic question recommendation for question answering communities. In *Proc. of WWW*, 2009.

[18] T. Sakai, D. Ishikawa, N. Kando, Y. Seki, K. Kuriyama, and C.-Y. Lin. Using graded-relevance metrics for evaluating community qa answer selection. In *Proc. of WSDM*, 2011.

[19] C. Shah and J. Pomerantz. Evaluating and predicting answer quality in community QA. In *Proc. of SIGIR*, 2010.

[20] M. A. Suryanto, E. P. Lim, A. Sun, and R. H. L. Chiang. Quality-aware collaborative question answering: methods and evaluation. In *Proc. of WSDM*, 2009.

[21] X.-J. Wang, X. Tu, D. Feng, and L. Zhang. Ranking community answers by modeling question-answer relationships via analogical reasoning. In *Proc. of SIGIR*, 2009.

[22] X. Xue, J. Jeon, and W. B. Croft. Retrieval models for question and answer archives. In *Proc. of SIGIR*, 2008.

[23] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In *Proc. of NIPS*, pages 321–328. MIT Press, 2004.

[24] T. C. Zhou, C.-Y. Lin, I. King, M. R. Lyu, Y.-I. Song, and Y. Cao. Learning to suggest questions in online forums. In *AAAI*. AAAI Press, 2011.

[25] X. Zhu. *Semi-Supervised Learning with Graphs*. PhD thesis, Carnegie Mellon University, 2005.

[26] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proc. of ICML*, 2003.