Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media

Towards Predicting the Best Answers in Community-Based Question-Answering Services

Qiongjie Tian and **Peng Zhang** and **Baoxin Li**
Computer Science and Engineering
Arizona State University
Tempe, Arizona, 85281
{qtian5,pzhang41,baoxin.li}@asu.edu

Abstract

Community-based question-answering (CQA) services contribute to solving many difficult questions we have. For each question in such services, one best answer can be designated, among all answers, often by the asker. However, many questions on typical CQA sites are left without a best answer even if when good candidates are available. In this paper, we attempt to address the problem of predicting if an answer may be selected as the best answer, based on learning from labeled data. The key tasks include designing features measuring important aspects of an answer and identifying the most importance features. Experiments with a Stack Overflow dataset show that the contextual information among the answers should be the most important factor to consider.

Introduction

Community-based question-answering (CQA) services help people solve many difficult questions. The importance and huge societal impact of such services are evidenced by the heavy traffic observed on popular CQA sites like Yahoo Answers (answers.yahoo.com), Baidu Zhidao (zhidao.baidu.com), and Stack Overflow (stackoverflow.com). On a CQA site, a person (the asker) posts a question and waits for answers from other users (the answerers). If multiple answers are provided, the asker can select the most suitable one, which is called the *accepted answer* or the *best answer*. Questions that do not have a designated best answer are stamped as "not-answered". Not every asker always selects the best answer for his/her question. This could be simply due to lack of action, or due to the difficulties in deciding on the best answer. As a result, many questions are left as "not-answered" (e.g., see (Yang et al. 2011)). Not-answered questions do not facilitate knowledge exchange, as other users would hesitate to rely on them for information, given their "not-answered" labels, even if in reality there may be many good candidate answers posted. Some sites also delete such not-answered questions after certain time of their posting, resulting in lost knowledge if there is indeed a suitable answer posted already. Towards addressing these problems, this paper focuses on learning from labeled data to predict whether an answer should be selected as the best answer. The study on best answer prediction can also contribute to the understanding of answer quality and help users improve their answers.

Copyright © 2013, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

For a candidate answer A_c to be considered as the best answer, in general three factors need to be assessed: (1) the quality of the answer content (e.g., its readability); (2) whether the answer contributes to solving the given question Q; and (3) how it competes with other answers A_i. These are schematically illustrated in Figure 1). We call the third factor *contextual information* since it is *relative* in nature. While there have been some reported studies ((Adamic et al. 2008; Shah and Pomerantz 2010; Blooma, Chua, and Goh 2010), to be detailed in the next section) on predicting the best answer, it remains to be fully explored to consider all these factors coherently and to evaluate the importance of the contextual information in solving the problem. This is the objective of this study.

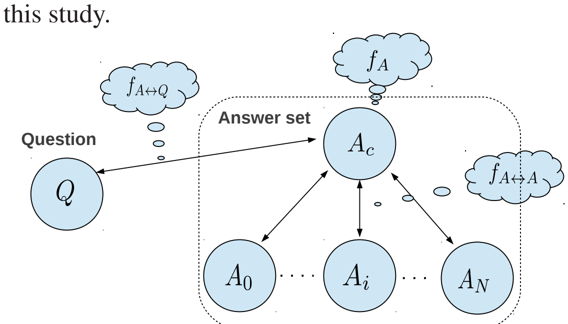

Figure 1: It illustrates three factors in assessing the likelihood of an answer A_c under consideration as the best answer: the dash-lined rectangle indicates the answer set to the question Q. $f_{A\leftrightarrow Q}$ is the set of features measuring relevance of A_c to Q, f_A is the set of features measuring the inherent quality of A_c, and $f_{A\leftrightarrow A}$ is the set of features measuring the competition between A_c and the other answers $A_0, \cdots, A_N$.

The major contribution of the work is twofold. Firstly, based on the analysis of a large CQA dataset, we designed features to measure the three key factors in selecting the best answer, especially contextual information. Secondly, through designing and evaluating a learning approach using these features to predict whether an answer may be selected as the best answer, we studied the importance of the factors based on their contribution to making the correct prediction.

Related Work

There are a few related studies in the literature. Liu *et al.* worked on predicting the asker's satisfaction with the answers (Liu, Bian, and Agichtein 2008). The features used do not measure contextual information among the answers. Harper *et al.* studied answer quality by answering two research questions: how the answer quality in different CQA sites is different from each other and how askers receive better answers (Harper et al. 2008). They found that fee-based CQA sites are more likely to receive high quality answers. Jeon *et al.* continued to work on the further effect of price on answer quality in fee-based CQA sites (Jeon, Kim, and Chen 2010). For the answer quality in different CQA sites, Fichman also made a detailed comparison (Fichman 2011). Shah *et al.* worked on the best answer prediction (Shah and Pomerantz 2010). In their work, they extracted features which contain information from the questions, the answers, and the users. But there is no consideration on the relationship between the answers and the questions, or relationship among the answers. This is the same case with the work in (Blooma, Chua, and Goh 2010). Yang *et al.* worked on predicting whether a question will receive the best answer and analyzed which factors contribute to solving the problem (Yang et al. 2011). Adamic *et al.* studied activity characteristics and mentioned how to predict whether one answer is the best answer given the question with its answers (Adamic et al. 2008), using content feature proposed in (Agichtein et al. 2008). In both cases, not all the factors were considered and especially the contextual information among the answers was not explicitly employed.

Stack Overflow Description

This study is based on Stack Overflow, a CQA site for computer programming, which was selected for its good quality control on the questions (and accordingly the answers) since any post unrelated to programming will be deleted automatically or via voting by senior users. Each question has three main parts: *title*, *body* and *tags*. In the body part, askers can describe their problems in detail. They may use figures or URL links etc. For tags, they may choose at most five existing terms that are most related to the question, or they can create new tags. Each question may receive multiple answers. For each question or answer, users can add *comments* to further discuss it. If one comment is good for solving the problem, it will be awarded with a *score* which shows in front of the comment. For each post (a question or an answer), it will have *upvotes* or *downvotes* from senior users and the corresponding askers or answerers will earn or lose reputation correspondingly. For a question, after it receives multiple answers, the asker can select one which in his or her opinion is most suitable for his or her question. The selected answer is called *Accepted Answer*, which is used in this study interchangeably as the best answer. Figure 2 illustrates one sample on Stack Overflow.

The dataset we used in this paper was downloaded from Stack Overflow for questions and answers posted before August 2012. The original dataset has contains 3,453,742 questions and 6,858,133 answers. In our experiment, we first select questions posted in June 2011 and then track all the answers or comments until August 2012. That is, each question was posted for more than one full years before the answers were collected. In this way, we may assume that all the questions were given enough time to gather good answers. This resulted in a subset of 103,793 questions and 196,145 answers, on which the later experimental results were based.

Figure 2: This is a sample to show the questions and answers on Stack Overflow site.

Features Description

As described above, our goal is to predict whether an answer will be selected as the best answer. We now design features for a given answer (with its corresponding question and other answers). The questions and answers are first preprocessed via standard procedures as illustrated in Figure 3, where the original text streams (sentences) are represented by the vector-space unigram model with TF-IDF weights (Shtok et al. 2012). In subsequent discussion, this pre-process result will contribute to the extraction of the following features (Table 1), corresponding to the three factors (Figure 1) discussed previously.

Answer Context: $f_{A \leftrightarrow A}$

To describe the context information, we use three features: similarity between the answer A_c under consideration and other answers A_i to the same question, the number of A_i, and the order A_c was created ans_index (e.g. by sorting the creation time, we know that A_c is the 4th answer to its question). The similarity feature has three dimensions: average, minimum and maximum similarity between A_c and A_i as defined below:

$$ave\_Ans\_sim = \frac{\sum_{i \neq c} sim(A_c, A_i)}{num(A_{i \neq c})} \tag{1}$$

$$min\_Ans\_sim = \min_{i \neq c} sim(A_c, A_i) \tag{2}$$

$$max\_Ans\_sim = \max_{i \neq c} sim(A_c, A_i) \tag{3}$$

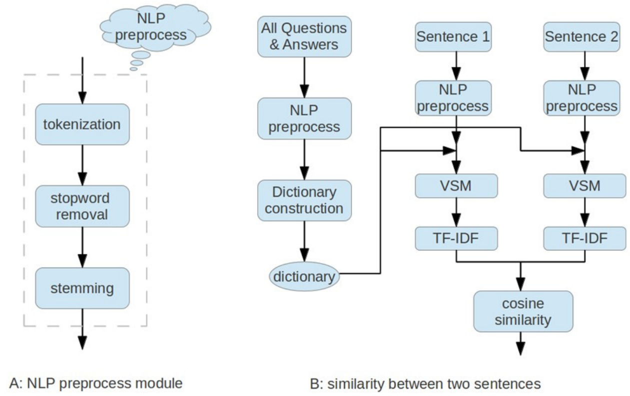

Figure 3: This figure shows the process to compute the similarity between two sentences. Part A is the pre-process module which is used in Part B. Part B is the flow chart to show how to compute the similarity.

where $sim(\cdot, \cdot)$ is the cosine similarity as in Figure 3 and $num(A_{i \neq c})$ is the total number of other answers A_i.

Question-Answer relationship: $f_{A \leftrightarrow Q}$

This group of features are based on the similarity between A_c and Q, which is $sim(A_c, Q)$, and also the time lag between the postings of the question and the answer, which is $timeSlot(A_c, Q)$. Since each question consists of a title and a body, to compute the similarity, we combine the title and the body before calculating the cosine similarity. Because the question can receive an answer at any time if it is not locked or closed, the time lapse between question and answer varies dramatically (e.g., from a few seconds to one year in our data). Thus, we represent this lag using logarithm scale.

$$QA\_sim = sim(A_c, Q) \tag{4}$$

$$timeSlot = timeSlot(A_c, Q) \tag{5}$$

Answer content: f_A

To describe the content quality of an answer, multiple features are defined below:

- Features from the answer body: the length of answer body, whether it has illustration pictures/codes, whether it refers to other web pages using URL, etc. Moreover, if one answer has a clear paragraph structure instead of messing everything up into one paragraph, it will be easy to read and then likely to be selected as a best answer. Thus, the readability of the answer also affects whether the answer will be selected as best answer and we define it as features related with paragraph length (Eq.6).

$$readability = [\max_i(L_i), \frac{1}{M}\sum_{i=1}^{M} L_i] \tag{6}$$

  where L_i is the length of i_{th} paragraph of the answer and M is the total number of paragraphs.

- Features from an answer's comments: The features are the number and average score of the comments and the variance of the scores.

Table 1 summarizes the above three types of features. Together, we compute a 16-dimensional feature vector for a candidate answer under consideration.

Prediction via Classification

With the features extracted for a candidate answer, we predict if it may be selected as the best answer through learning a classifier using labelled data: feature vectors corresponding to best answers and non-best-answers according to the ground-truth are used to learn a 2-class classifier. The classifier we used is based on the random forest algorithm(Breiman 2001). Random forest is an efficient algorithm to classify large dataset. It also provides an efficient approach to computing feature importance, which is useful for us to analyze the importance of each feature Table 1.

Table 1: Features designed for an answer A_c to a question Q. A_i are other answers to Q.

group	index	symbol	feature description
f_A	0,1	*ave_comment*, *var_comment*	they are the average and variance of the scores of the comments to A_c.
	2	*comment_num*	A_c's comments number.
	3, 4, 5	*URL_tag*, *pic*, *code*	they show whether A_c has a URL tag, illustration figures, or codes.
	6	*ans_len*	it is the length of A_c.
	7, 8	*readability*	they show whether A_c is easy to read, see Eq.6
$f_{A \leftrightarrow Q}$	9	*QA_sim*	the similarity between A_c and Q. (Figure 3).
	10	*timeSlot*	the difference between A_c's creation time and Q's.
$f_{A \leftrightarrow A}$	11, 12, 13	*ave_Ans_sim*, *min_Ans_sim*, *max_Ans_sim*	the average, minimum, maximum of similarities between A_c and A_i.
	14	*competitor_num*	the number of A_i.
	15	*ans_index*	the order that A_c was created. E.g. it is the 2_{nd} answer to the question.

Experimental Results

The experiments were based on the Stack Overflow dataset described earlier. Among the 103,793 questions and 196,145 answers used, there are 4,950 questions that do not have any answer and 45,715 questions with only one answers. For questions with only one, 16,986 of them have no best answers while 28,729 having the best answers. We used all 196,145 answers in our experiment, with the best answers as positive samples and the negative samples being the answers that are not best answers.

We use random forest classifier to do classification and twofold cross-validation. The average accuracy is shown in Table 2. We emphasize that the focus of this study is on analyzing only features extracted from the questions and answers without using user-specific information. User-specific

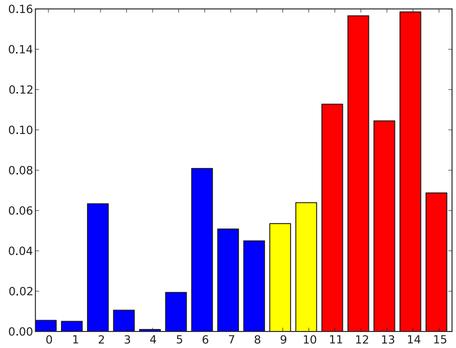

Figure 4: The distribution of feature importances. The bars correspond to 16 features defined in Table 1, respectively.

information, when available, can be used to further improve the performance as done in (Yang *et al.* 2011).

The distribution of the feature importance is shown in Figure 4. Both Figure 4 and Table 2 indicate that features from the answer context $f_{A \leftrightarrow A}$ contribute the most. We also compute the average feature importance from the three groups of features. For features from the answer context, the average feature importance is 0.1202. For the features from the question-answer relationship, the average feature importance is 0.05871. For the features from the answer content, the average feature importance is 0.03128. This also shows the importance of $f_{A \leftrightarrow A}$. In the following, we discuss feature importances based on Figure 4, respectively.

Table 2: Prediction accuracy for different feature groups. $f_{A \leftrightarrow A}$, $f_{A \leftrightarrow Q}$, f_A are three groups of features we described in the previous sections.

Features	$f_{A \leftrightarrow A}$	$f_{A \leftrightarrow Q}$	f_A	all
Accuracy	70.71%	60.27%	65.59%	72.27%

In the group $f_{A \leftrightarrow A}$, the most important feature is *competitor_num*. This suggests that the more competitors the answer A_c has, the less likely is may be selected as the best answer. The feature *min_Ans_sim* has slightly less but comparable importance as *competitor_num*. This shows that the best answer is usually most different from the others. However it does not mean the best answer and the competitors should be totally different. Since all the answers aim at answering the same questions, they also should have similarity. We can see this from the importance of ave_Ans_sim.

In the group $f_{A \leftrightarrow Q}$, the feature $timeSlot$ contributes more than the feature QA_sim. This shows that earlier answers have a higher chance to be selected as the best answer.

Within the group f_A, $comment\_num$ and ans_len contribute more than the others. This suggests that the best answer is usually the one with more details and comments. This is reasonable and intuitive. The $readability$ feature also contributes significantly, suggesting that answers that are easy to read are likely to be selected.

Conclusion and Future work

We studied the problem of predicting the best answer on CQA sites. Our experiments and analysis with a reasonably large dataset have shown that some features, and in particular those reflecting the contextual information among the answers, are more important for the task. The results also suggest that the features designed in the paper appear to be able to do the job reasonably well. In the future, we plan to study the importance of user-centric information (e.g., usage history, location etc.) for the prediction problem.

Acknowledgments

This work was supported in part by a grant (#1135616) from the National Science Foundation. Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Adamic, L.; Zhang, J.; Bakshy, E.; and Ackerman, M. 2008. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, 665–674. ACM.

Agichtein, E.; Castillo, C.; Donato, D.; Gionis, A.; and Mishne, G. 2008. Finding high-quality content in social media. In *Proceedings of the international conference on Web search and web data mining*, 183–194. ACM.

Blooma, M. J.; Chua, A.-K.; and Goh, D.-L. 2010. Selection of the best answer in cqa services. In *Information Technology: New Generations (ITNG), 2010 Seventh International Conference on*, 534–539. IEEE.

Breiman, L. 2001. Random forests. *Machine learning* 45(1):5–32.

Fichman, P. 2011. A comparative assessment of answer quality on four question answering sites. *Journal of Information Science* 37(5):476–486.

Harper, F. M.; Raban, D.; Rafaeli, S.; and Konstan, J. A. 2008. Predictors of answer quality in online q&a sites. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, 865–874. ACM.

Jeon, G. Y.; Kim, Y.-M.; and Chen, Y. 2010. Re-examining price as a predictor of answer quality in an online q&a site. In *Proceedings of the 28th international conference on Human factors in computing systems*, 325–328. ACM.

Liu, Y.; Bian, J.; and Agichtein, E. 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*, 483–490. ACM.

Shah, C., and Pomerantz, J. 2010. Evaluating and predicting answer quality in community qa. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 411–418. Citeseer.

Shtok, A.; Dror, G.; Maarek, Y.; and Szpektor, I. 2012. Learning from the past: answering new questions with past answers. In *Proceedings of the 21st international conference on World Wide Web*, 759–768. ACM.

Yang, L.; Bao, S.; Lin, Q.; Wu, X.; Han, D.; Su, Z.; and Yu, Y. 2011. Analyzing and predicting not-answered questions in community-based question answering services. In *Proceedings of AAAI*, 1273–1278.