Question Difficulty Estimation in Community Question Answering Services*

Jing Liu[†] **Quan Wang**[‡] **Chin-Yew Lin**[♯] **Hsiao-Wuen Hon**[♯]
[†]Harbin Institute of Technology, Harbin 150001, P.R.China
[‡]Peking University, Beijing 100871, P.R.China
[♯]Microsoft Research Asia, Beijing 100080, P.R.China
jliu@ir.hit.edu.cn quanwang1012@gmail.com {cyl,hon}@microsoft.com

Abstract

In this paper, we address the problem of estimating question difficulty in community question answering services. We propose a competition-based model for estimating question difficulty by leveraging pairwise comparisons between questions and users. Our experimental results show that our model significantly outperforms a PageRank-based approach. Most importantly, our analysis shows that the text of question descriptions reflects the question difficulty. This implies the possibility of predicting question difficulty from the text of question descriptions.

1 Introduction

In recent years, community question answering (CQA) services such as Stackoverflow[1] and Yahoo! Answers[2] have seen rapid growth. A great deal of research effort has been conducted on CQA, including: (1) question search (Xue et al., 2008; Duan et al., 2008; Suryanto et al., 2009; Zhou et al., 2011; Cao et al., 2010; Zhang et al., 2012; Ji et al., 2012); (2) answer quality estimation (Jeon et al., 2006; Agichtein et al., 2008; Bian et al., 2009; Liu et al., 2008); (3) user expertise estimation (Jurczyk and Agichtein, 2007; Zhang et al., 2007; Bouguessa et al., 2008; Pal and Konstan, 2010; Liu et al., 2011); and (4) question routing (Zhou et al., 2009; Li and King, 2010; Li et al., 2011).

However, less attention has been paid to ***question difficulty estimation*** in CQA. Question difficulty estimation can benefit many applications: (1) Experts are usually under time constraints. We do not want to bore experts by routing every question (including both easy and hard ones) to them. Assigning questions to experts by matching question difficulty with expertise level, not just question topic, will make better use of the experts' time and expertise (Ackerman and McDonald, 1996). (2) Nam et al. (2009) found that winning the point awards offered by the reputation system is a driving factor in user participation in CQA. Question difficulty estimation would be helpful in designing a better incentive mechanism by assigning higher point awards to more difficult questions. (3) Question difficulty estimation can help analyze user behavior in CQA, since users may make strategic choices when encountering questions of different difficulty levels.

To the best of our knowledge, not much research has been conducted on the problem of estimating question difficulty in CQA. The most relevant work is a PageRank-based approach proposed by Yang et al. (2008) to estimate task difficulty in crowdsourcing contest services. Their key idea is to construct a graph of tasks: creating an edge from a task t_1 to a task t_2 when a user u wins task t_1 but loses task t_2, implying that task t_2 is likely to be more difficult than task t_1. Then the standard PageRank algorithm is employed on the task graph to estimate PageRank score (i.e., difficulty score) of each task. This approach implicitly assumes that task difficulty is the only factor affecting the outcomes of competitions (i.e. the best answer). However, the outcomes of competitions depend on both the difficulty levels of tasks and the expertise levels of competitors (i.e.

*This work was done when Jing Liu and Quan Wang were visiting students at Microsoft Research Asia. Quan Wang is currently affiliated with Institute of Information Engineering, Chinese Academy of Sciences.

[1]http://stackoverflow.com

[2]http://answers.yahoo.com

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 85–90,
Seattle, Washington, USA, 18-21 October 2013.

other answerers).

Inspired by Liu et al. (2011), we propose a competition-based approach which jointly models question difficulty and user expertise level. Our approach is based on two intuitive assumptions: (1) given a question answering thread, the difficulty score of the question is higher than the expertise score of the asker, but lower than that of the best answerer; (2) the expertise score of the best answerer is higher than that of the asker as well as all other answerers. Given the two assumptions, we can determine the question difficulty score and user expertise score through pairwise comparisons between (1) a question and an asker, (2) a question and a best answerer, (3) a best answerer and an asker, and (4) a best answerer and all other non-best answerers.

The main contributions of this paper are:

• We propose a competition-based approach to estimate question difficulty (Sec. 2). Our model significantly outperforms the PageRank-based approach (Yang et al., 2008) for estimating question difficulty on the data of Stack Overflow (Sec. 3.2).

• Additionally, we calibrate question difficulty scores across two CQA services to verify the effectiveness of our model (Sec. 3.3).

• Most importantly, we demonstrate that different words or tags in the question descriptions indicate question difficulty levels. This implies the possibility of predicting question difficulty purely from the text of question descriptions (Sec. 3.4).

2 Competition based Question Difficulty Estimation

CQA is a virtual community where people can ask questions and seek opinions from others. Formally, when an asker u_a posts a question q, there will be several answerers to answer her question. One answer among the received ones will be selected as the best answer by the asker u_a or voted by the community. The user who provides the best answer is called the best answerer u_b, and we denote the set of all non-best answerers as $S = \{u_{o_1}, \cdots, u_{o_M}\}$. Assuming that question difficulty scores and user expertise scores are expressed on the same scale, we make the following two assumptions:

• The difficulty score of question q is higher than the expertise score of asker u_a, but lower than that of the best answerer u_b. This is intuitive since the best answer u_b correctly responds to question q that asker u_a does not know.

• The expertise score of the best answerer u_b is higher than that of asker u_a and all answerers in S. This is straightforward since the best answerer u_b solves question q better than asker u_a and all non-best answerers in S.

Let's view question q as a pseudo user u_q. Taking a competitive viewpoint, each pairwise comparison can be viewed as a two-player competition with one winner and one loser, including (1) one competition between pseudo user u_q and asker u_a, (2) one competition between pseudo user u_q and the best answerer u_b, (3) one competition between the best answerer u_b and asker u_a, and (4) $|S|$ competitions between the best answerer u_b and all non-best answers in S. Additionally, pseudo user u_q wins the first competition and the best answerer u_b wins all remaining $(|S| + 2)$ competitions.

Hence, the problem of estimating the question difficulty score (and the user expertise score) is cast as a problem of learning the relative skills of players from the win-loss results of the generated two-player competitions. Formally, let $\mathcal{Q}$ denote the set of all questions in one category (or topic), and $\mathcal{R}_q$ denote the set of all two-player competitions generated from question $q \in \mathcal{Q}$, i.e., $\mathcal{R}_q = \{(u_a \prec u_q), (u_q \prec u_b), (u_a \prec u_b), (u_{o_1} \prec u_b), \cdots, (u_{o_{|S|}} \prec u_b)\}$, where $j \prec i$ means that user i beats user j in the competition. Define

$$\mathcal{R} = \bigcup_{q \in \mathcal{Q}} \mathcal{R}_q \tag{1}$$

as the set of all two-player competitions. Our problem is then to learn the relative skills of players from $\mathcal{R}$. The learned skills of the pseudo question users are question difficulty scores, and the learned skills of all other users are their expertise scores.

TrueSkill In this paper, we follow (Liu et al., 2011) and apply TrueSkill to learn the relative skills of players from the set of generated competitions $\mathcal{R}$ (Equ. 1). TrueSkill (Herbrich et al., 2007) is a Bayesian skill rating model that is developed for estimating the relative skill levels of players in games. In this paper, we present a two-player version of TrueSkill with no-draw.

TrueSkill assumes that the practical performance of each player in a game follows a normal distribu-

tion $\mathcal{N}(\mu, \sigma^2)$, where μ means the skill level of the player and σ means the uncertainty of the estimated skill level. Basically, TrueSkill learns the skill levels of players by leveraging Bayes' theorem. Given the current estimated skill levels of two players (priori probability) and the outcome of a new game between them (likelihood), TrueSkill model updates its estimation of player skill levels (posterior probability). TrueSkill updates the skill level μ and the uncertainty σ intuitively: (a) if the outcome of a new competition is expected, i.e. the player with higher skill level wins the game, it will cause small updates in skill level μ and uncertainty σ; (b) if the outcome of a new competition is unexpected, i.e. the player with lower skill level wins the game, it will cause large updates in skill level μ and uncertainty σ. According to these intuitions, the equations to update the skill level μ and uncertainty σ are as follows:

$$\mu_{winner} = \mu_{winner} + \frac{\sigma^2_{winner}}{c} \cdot v\left(\frac{t}{c}, \frac{\varepsilon}{c}\right), \quad (2)$$

$$\mu_{loser} = \mu_{loser} - \frac{\sigma^2_{loser}}{c} \cdot v\left(\frac{t}{c}, \frac{\varepsilon}{c}\right), \quad (3)$$

$$\sigma^2_{winner} = \sigma^2_{winner} \cdot \left[1 - \frac{\sigma^2_{winner}}{c^2} \cdot w\left(\frac{t}{c}, \frac{\varepsilon}{c}\right)\right], \quad (4)$$

$$\sigma^2_{loser} = \sigma^2_{loser} \cdot \left[1 - \frac{\sigma^2_{loser}}{c^2} \cdot w\left(\frac{t}{c}, \frac{\varepsilon}{c}\right)\right], \quad (5)$$

where $t = \mu_{winner} - \mu_{loser}$ and $c^2 = 2\beta^2 + \sigma^2_{winner} + \sigma^2_{loser}$. Here, ε is a parameter representing the probability of a draw in one game, and $v(t, \varepsilon)$ and $w(t, \varepsilon)$ are weighting factors for skill level μ and standard deviation σ respectively. Please refer to (Herbrich et al., 2007) for more details. In this paper, we set the initial values of the skill level μ and the standard deviation σ of each player the same as the default values used in (Herbrich et al., 2007).

3 Experiments

3.1 Data Set

In this paper, we use Stack Overflow (SO) for our experiments. We obtained a publicly available data set[3] of SO between July 31, 2008 and August 1, 2012. SO contains questions with various topics, such as programming, mathematics, and English. In this paper, we use SO C++ programming (SO/CPP) and mathematics[4] (SO/Math) questions for our main experiments. Additionally, we use the data of Math Overflow[5] (MO) for calibrating question difficulty scores across communities (Sec. 3.3). The statistics of these data sets are shown in Table 1.

	SO/CPP	**SO/Math**	**MO**
# of questions	122,012	51,174	27,333
# of answers	357,632	94,488	65,966
# of users	67,819	16,961	12,064

Table 1: The statistics of the data sets.

To evaluate the effectiveness of our proposed model for estimating question difficulty scores, we randomly sampled 300 question pairs from both SO/CPP and SO/Math, and we asked experts to compare the difficulty of every pair. We had two graduate students majoring in computer science annotate the SO/CPP question pairs, and two graduate students majoring in mathematics annotate the SO/Math question pairs. When annotating each question pair, only the titles, descriptions, and tags of the questions were shown, and other information (e.g. users, answers, etc.) was excluded. Given each pair of questions (q_1 and q_2), the annotators were asked to give one of four labels: (1) $q_1 \succ q_2$, which means that the difficulty of q_1 was higher than q_2; (2) $q_1 \prec q_2$, which means that the difficulty of q_1 was lower than q_2; (3) $q_1 = q_2$, which means that the difficulty of q_1 was equal to q_2; (4) Unknown, which means that the annotator could not make a decision. The agreements between annotators on both SO/CPP (kappa value $= 0.741$) and SO/Math (kappa value $= 0.873$) were substantial. When evaluating models, we only kept the pairs that annotators had given the same labels. There were 260 SO/CPP question pairs and 280 SO/Math question pairs remaining.

3.2 Accuracy of Question Difficulty Estimation

We employ a standard evaluation metric for information retrieval: accuracy (Acc), defined as follows:

$$Acc = \frac{\text{the number of correct pairwise comparisons}}{\text{the total number of pairwise comparisons}}.$$

We use the *PageRank*-based approach proposed by Yang et al. (2008) as a baseline. As described in

[3] http://blog.stackoverflow.com/category/cc-wiki-dump/

[4] http://math.stackexchange.com

[5] http://mathoverflow.net

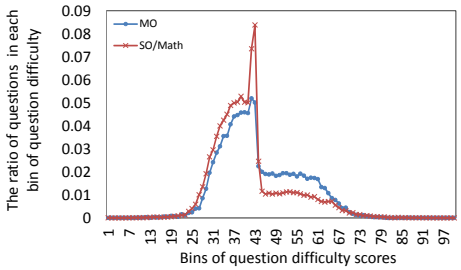


Figure 1: The distributions of calibrated question difficulty scores of MO and SO/Math.

Sec. 1, this is the most relevant method for our problem. Table 2 gives the accuracy of the baseline and our *Competition*-based approach on SO/CPP and SO/Math. From the results, we can see that (1) the proposed *Competition*-based approach significantly outperformed the *PageRank*-based approach on both data sets; (2) *PageRank*-based approach only achieved a similar performance as randomly guessing. This is because the *PageRank*-based approach only models the outcomes of competitions affected by question difficulty. However, the outcomes of competitions depend on both the question difficulty levels and the expertise levels of competitors. Our *Competition*-based approach considers both these factors for modeling the competitions. The experimental results demonstrate the advantage of our approach.

	Acc@**SO/CPP**	*Acc*@**SO/Math**
PageRank	50.38%	48.93%
Competition	66.54%	71.79%

Table 2: Accuracy on SO/CPP and SO/Math.

3.3 Calibrating Question Difficulty across CQA Services

Both MO and SO/Math are CQA services for asking mathematics questions. However, these two services are designed for different audiences, and they have different types of questions. MO's primary goal is asking and answering research level mathematics questions[6]. In contrast, SO/Math is for people studying mathematics at any level in related fields[7]. Usually, the community members in MO are not interested in basic mathematics questions. If a posted question is too elementary, someone will suggest moving it to SO/Math. Similarly, if a posted question is advanced, the community members in SO/Math will recommend moving it to MO. Hence, it is expected that the ratio of difficult questions in MO is higher than SO/Math. In this section, we examine whether our competition-based model can identify such differences.

We first calibrate the estimated question difficulty scores across these two services on a same scale. The key idea is to link the users who participate in both services. In both MO and SO/Math, users can specify their home pages. We assume that if a user u_1 on MO and a user u_2 on SO/Math have the same home page URL, they should be linked as one natural person in the real world. We successfully linked 633 users. They provided $18,196$ answers in SO/Math among which $10,993$ (60.41%) were selected as the best answers. In contrast, they provided $8,044$ answers in MO among which $3,215$ (39.97%) were selected as the best answers. This shows that these users reflect more competitive contests in MO. After the common users are linked, we have a joint data set of MO and SO/Math. Then, we can calibrate the estimated question difficulty scores across the two services by performing the competition-based model on the joint data set. Figure 1 shows the distributions of the calibrated question difficulty scores of MO and SO/Math on the same scale. As expected, we observed that the ratio of difficult questions in MO was higher than SO/Math. Additionally, these two distributions were significantly different (Kolmogorov-Smirnov Test, p-value < 0.05). This demonstrates that our competition-based model successfully identified the difference between questions on two CQA services.

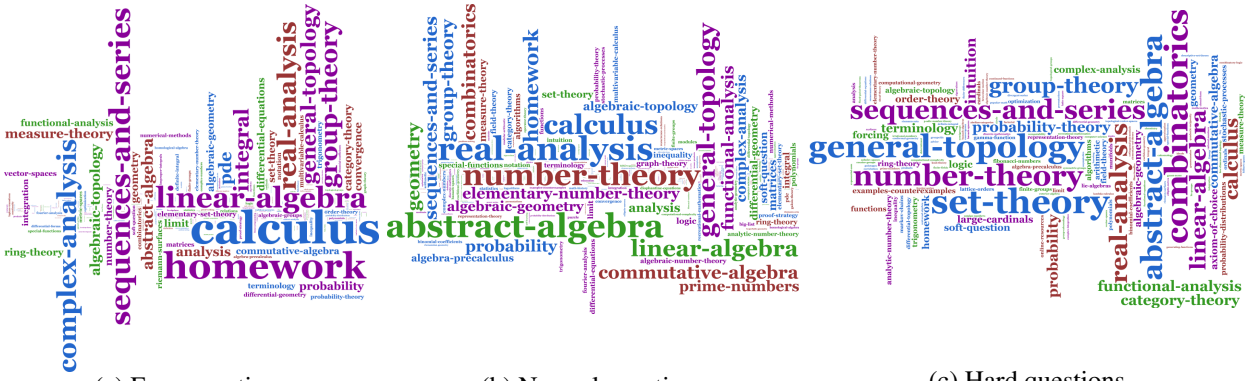
3.4 Analysis on the Question Descriptions

In this section, we analyze the text of question descriptions on the scale of question difficulty scores estimated by the competition model.

Micro Level We first examine the frequency distributions of individual words over the question difficulty scores. Figure 3 shows the examples of four words in SO/CPP. We observe that the words *'list'* and *'array'* have the lowest mean of difficulty scores, compared to the words *'virtual'* and *'gcc'*. This is reasonable, since *'list'* and *'array'* are related

[6] http://mathoverflow.net/faq

[7] http://area51.stackexchange.com/proposals/3355/mathematics



(a) Easy questions (b) Normal questions (c) Hard questions

Figure 2: Tag clouds on SO/Math questions with different difficulty levels

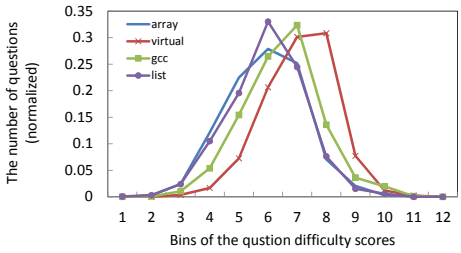


Figure 3: The frequency distributions of words on the scale of question difficulty scores (SO/CPP).

to basic concepts in programming language, while *'virtual'* and *'gcc'* are related to more advanced topics. It can be observed that the order of the means of the difficulty scores of these words are well aligned to our learning process.

Macro Level We evenly split the range of question difficulty scores into n buckets, and we grouped the questions into the n buckets according to which bucket their difficulty scores were in. Then, we had n question buckets and each bucket corresponded to a word distribution of questions. Let variable X denote the distance between the difficulty scores in two question buckets (which is the difference between the average difficulty scores of questions in the two buckets), and variable Y denote the Jensen-Shannon distance between word distributions in two question buckets. We examined the correlation between variable X and variable Y. The experimental results showed that the correlation between these two variables were strongly positive. Specifically, the correlation coefficient on SO/CPP was 0.8129 and on SO/Math was 0.7412. In other words, when the distance between the difficulty scores of two buckets become larger, the two word distributions in the two buckets become less similar, and vice versa.

We further visualized the word distribution in each question bucket. We set n as 3, and we had three question buckets: (1) easy questions; (2) normal questions; and (3) hard questions. Figure 3.4 plots the tag clouds of SO/Math questions in the three buckets. The size of tags is proportional to the frequency of tags in each bucket. We observed that (1) the tag *'homework'* and *'calculus'* become smaller from easy questions to hard questions; (2) the tag *'set-theory'* becomes larger. These observations also reflect our learning process.

The above experimental results show that different words or tags of question descriptions reflect the question difficulty levels. This implies the possibility of predicting question difficulty purely from the text of question descriptions.

4 Conclusion and Future Work

In this paper, we address the problem of estimating question difficulty in CQA services. Our proposed competition-based model for estimating question difficulty significantly outperforms the PageRank-based approach. Most importantly, our analysis shows that the text of question descriptions reflects the question difficulty. In the future, we would like to explore predicting question difficulty from the text of question descriptions. We also will investigate non-technical areas, where there might be no strongly distinct notion of experts and non-experts.

Acknowledgments

We would like to thank Yunbo Cao and Jie Cai for their valuable suggestions for this paper, and the anonymous reviewers for their helpful comments.

References

M.S. Ackerman and D.W. McDonald. 1996. Answer garden 2: merging organizational memory with collaborative help. In *Proceedings of CSCW*.

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

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

M. Bouguessa, B. Dumoulin, and S. Wang. 2008. Identifying authoritative actors in question-answering forums: the case of yahoo! answers. In *Proceeding of SIGKDD*.

Xin Cao, Gao Cong, Bin Cui, and Christian S Jensen. 2010. A generalized framework of exploring category information for question retrieval in community question answer archives. In *Proceedings of WWW*.

H. Duan, Y. Cao, C.Y. Lin, and Y. Yu. 2008. Searching questions by identifying question topic and question focus. In *Proceedings of ACL*.

R. Herbrich, T. Minka, and T. Graepel. 2007. Trueskill: A bayesian skill rating system. In *Proceedings of NIPS*.

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

Zongcheng Ji, Fei Xu, Bin Wang, and Ben He. 2012. Question-answer topic model for question retrieval in community question answering. In *Proceedings of CIKM*.

P. Jurczyk and E. Agichtein. 2007. Discovering authorities in question answer communities by using link analysis. In *Proceedings of CIKM*.

B. Li and I. King. 2010. Routing questions to appropriate answerers in community question answering services. In *Proceedings of CIKM*.

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

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

J. Liu, Y.I. Song, and C.Y. Lin. 2011. Competition-based user expertise score estimation. In *Proceedings of SIGIR*.

K.K. Nam, M.S. Ackerman, and L.A. Adamic. 2009. Questions in, knowledge in?: a study of naver's question answering community. In *Proceedings of CHI*.

A. Pal and J.A. Konstan. 2010. Expert identification in community question answering: exploring question selection bias. In *Proceedings of CIKM*.

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

Xiaobing Xue, Jiwoon Jeon, and W Bruce Croft. 2008. Retrieval models for question and answer archives. In *Proceedings of SIGIR*.

Jiang Yang, Lada Adamic, and Mark Ackerman. 2008. Competing to share expertise: the taskcn knowledge sharing community. In *Proceedings of ICWSM*.

J. Zhang, M.S. Ackerman, and L. Adamic. 2007. Expertise networks in online communities: structure and algorithms. In *Proceedings of WWW*.

Weinan Zhang, Zhaoyan Ming, Yu Zhang, Liqiang Nie, Ting Liu, and Tat-Seng Chua. 2012. The use of dependency relation graph to enhance the term weighting in question retrieval. In *Proceedings of COLING*.

Y. Zhou, G. Cong, B. Cui, C.S. Jensen, and J. Yao. 2009. Routing questions to the right users in online communities. In *Proceedings of ICDE*.

Guangyou Zhou, Li Cai, Jun Zhao, and Kang Liu. 2011. Phrase-based translation model for question retrieval in community question answer archives. In *Proceedings of ACL*.