Information Sciences 302 (2015) 70–82

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Detecting high-quality posts in community question answering sites

Yuan Yao [a], Hanghang Tong [b], Tao Xie [c], Leman Akoglu [d], Feng Xu [a,*], Jian Lu [a]

[a] State Key Laboratory for Novel Software Technology, Nanjing University, China
[b] Arizona State University, USA
[c] University of Illinois at Urbana-Champaign, USA
[d] Stony Brook University, USA

ARTICLE INFO

Article history:
Received 3 August 2014
Received in revised form 15 November 2014
Accepted 9 December 2014
Available online 12 January 2015

Keywords:
CQA
Question and answer
Voting correlation
Voting prediction

ABSTRACT

Community question answering (CQA) has become a new paradigm for seeking and sharing information. In CQA sites, users can ask and answer questions, and provide feedback (e.g., by voting or commenting) to these questions/answers. In this article, we propose the early detection of high-quality CQA questions/answers. Such detection can help discover a high-impact question that would be widely recognized by the users in these CQA sites, as well as identify a useful answer that would gain much positive feedback from site users. In particular, we view the post quality from the perspective of the voting outcome. First, our key intuition is that the voting score of an answer is strongly positively correlated with that of its question, and we verify such correlation in two real CQA data sets. Second, armed with the verified correlation, we propose a family of algorithms to jointly detecting the high-quality questions and answers soon after they are posted in the CQA sites. We conduct extensive experimental evaluations to demonstrate the effectiveness and efficiency of our approaches. Overall, our algorithms can outperform the best competitor in prediction performance, while enjoying linear scalability with respect to the total number of posts.

1. Introduction

Community question answering (CQA) has become a new paradigm for seeking and sharing information. For example, millions of users now use CQA sites to search for solutions for their problems [15,19]. Example CQA sites include those general ones such as Yahoo! Answers[1] and Baidu Knows,[2] and those domain-specific ones like Stack Overflow[3] and Mathematics Stack Exchange.[4]

One major difference between CQA and traditional QA is from the volunteer efforts of site users. In addition to posting questions/answers, most of the existing CQA sites allow the site users to vote (e.g., upvote and downvote in Stack Overflow)

* Corresponding author.
E-mail addresses: yyao@smail.nju.edu.cn (Y. Yao), hanghang.tong@asu.edu (H. Tong), taoxie@illinois.edu (T. Xie), leman@cs.stonybrook.edu (L. Akoglu), xf@nju.edu.cn (F. Xu), lj@nju.edu.cn (J. Lu).

[1] http://answers.yahoo.com/.
[2] http://zhidao.baidu.com/.
[3] http://stackoverflow.com/.
[4] http://math.stackexchange.com/.

http://dx.doi.org/10.1016/j.ins.2014.12.038
0020-0255/

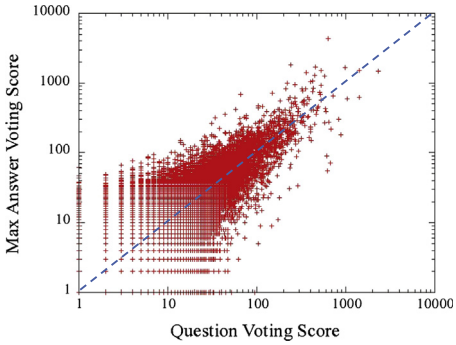

Fig. 1. The strong voting correlation between questions and their best answers in *SO* dataset. Pearson correlation coefficient $r = 0.6665$ with p-value < 0.0001. See Fig. 2 for more results.

for these questions/answers. On one hand, the voting mechanism as well as its reputation system provides the main incentives for tight involvement and productive competition of the whole community. On the other hand, the outcome of such voting, e.g., the difference between the number of the upvotes and downvotes that a question/answer receives from the site users (referred to as 'voting score'), provides a good indicator of the intrinsic value of a question/answer. To some extent, the voting score of a question/answer resembles the number of the citations that a research paper receives in the scientific publication domain. It reflects the net number of users who have a positive attitude toward the paper.

In this article, we view the post quality from the perspective of the voting outcome, and propose the early detection of high-quality CQA questions/answers. To date, a lot of efforts have been made to study the quality prediction problem in CQA sites. However, most of them treat questions and answers separately (see Section 6 for a review).

We conjecture that there exists *correlation* between the voting score of a question and that of its associated answer. Intuitively, an interesting question might obtain more attention from potential answerers and thus has a better chance to receive high-score answers. On the other hand, it might be very difficult for a low-score question to attract a high-score answer due to, e.g., its poor expression in language, or lack of interestingness in topic. Starting from this conjecture, we study two real CQA sites, i.e., Stack Overflow (*SO*), and Mathematics Stack Exchange (*Math*). Our key finding is that the voting score of an answer is indeed strongly positively correlated with that of its question (see Fig. 1). Such correlation structure consistently exists on both sites.

Armed with this observation, we propose a family of co-prediction algorithms (*CoPs*) to *jointly* predict the voting scores of questions and answers. In particular, we aim at identifying the potentially high-score posts soon after they are posted in the CQA sites. We conduct extensive experimental evaluations to demonstrate the effectiveness and efficiency of our approaches. Overall, our *joint* prediction approaches achieve up to 15.2% net precision improvement for answer prediction over the best competitor in one of the data sets we studied. In addition, the proposed *CoPs* algorithms enable us to predict the voting outcome of an answer *before* it actually appears on the site. Finally, our approaches scale linearly wrt the total number of questions and answers.

The main contributions of this paper include:

- *Empirical Findings*. We empirically study the voting scores of posts in CQA sites. To the best of our knowledge, we are the first to *quantitatively* validate the correlation between the voting scores of questions and those of their associated answers on two independent real data sets.
- *Algorithms and Evaluations*. We propose a family of co-prediction algorithms (*CoPs*) to jointly predict the voting scores of questions and answers. We further perform extensive experimental evaluations on two real data sets to demonstrate the effectiveness and efficiency of our approaches.

The rest of the paper is organized as follows. Section 2 presents the empirical studies about the voting scores of questions/answers. Sections 3 and 4 present the problem definitions and the proposed algorithms for the joint voting prediction problem, respectively. Section 5 presents the experimental results. Section 6 reviews related work, and Section 7 concludes the paper.

2. Empirical study

In this section, we perform an empirical study of the voting scores of questions and answers in *SO* (Stack Overflow) and *Math* (Mathematics Stack Exchange) data sets. They are popular CQA sites for programming and math, respectively. For both data sets, they are officially published and publicly available.[5] The statistics of the two data sets are summarized in Table 1.

We first study the overall correlation between the voting scores of questions and those of their answers. For a given question, there might be multiple answers. Thus, we report both the highest (i.e., the best answer) and the average voting scores

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