On early detection of high voted Q&A on Stack Overflow

Mahmood Neshati
Faculty of Computer Science and Engineering, Shahid Beheshti University, G.C, Tehran, Iran

A R T I C L E   I N F O
Article history:
Received 4 October 2016
Revised 12 January 2017
Accepted 7 February 2017

A B S T R A C T

Early detection of high quality content on community question answering platforms is an important emerging problem in which the main goal is the detection of high quality questions and answers in a short time right after their submission. Improving the process of question routing, reducing the number of questions with no answers, improving the user experience and also promoting the content quality of a CQA by rejecting low quality contents are all benefits of solving the early detection of high quality content problem in CQA. The main challenge of solving this problem is that the value of a few features is available in a short time after submission of a content in CQA. In other words, unlike previous related research, it is not possible to utilize comprehensive set of features to detect high quality content. In this paper, we view the content quality from the perspective of the voting outcome. Specifically, we consider those Q&A which will get more votes than a certain threshold as high quality posts. Analyzing large amount of data in a CQA, we observed two important patterns which help us with early detection of high quality content. We named the first pattern as accepted answer effect and the second pattern as answer competition effect. According to the first pattern, the chance of a high quality question to get an accepted answer is higher than the chance of other questions and vice versa. According to the second pattern, only few number of answers of a specific question will be high quality answers. We show that these patterns are valid in a short time after the submission of content on CQA. Utilizing these patterns, we propose a unified relational classification framework to solve the problem. In our proposed framework, the quality of a given question and its associated answers can be predicted simultaneously soon after their submission. We conduct several experiments on six data collections gathered from Stack Overflow in order to show the efficiency of the proposed models. Our experiments indicate that the performance of high quality content detection can improve up to 10.7% and 35.3% in comparison with a state-of-the-art independent classifier for questions and answers, respectively. Moreover, we found 1.2% and 11.8% F-measure gain in average versus a recent strong baseline by Yao et al. (2015) for questions and answers, respectively.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Community Question and Answering (CQA) websites are new data sharing platforms which help users to find the best answer to their information need directly, instead of searching and reading long and detailed documents. Successful CQA
websites include both general websites such as Yahoo! Answers\textsuperscript{1} and Quora\textsuperscript{2}, and also domain specific QA websites like StackOverflow\textsuperscript{3} and Mathematics Stack Exchange\textsuperscript{4}.

Previous research confirms that altruism can be considered as the main reason to motivate users to contribute in a community like a CQA (Ostrom, 1990). However, in addition, CQA websites usually use approaches like gamification to enhance the motivation of the users and accordingly to improve their quality and quantity of contribution. Voting on questions and answers, approving the satisfying answer (i.e. the accepted answer), awarding badges for an outstanding contribution of users, uniquely representing users by avatar, demonstrating the reputation of users and etc. are all elements of such gamification mechanism to enhance the motivation of users to provide high quality contents.

Some of these gamification elements (e.g. the number of votes of a question or answer, the reputation score of a user) can be considered as the quality indicators in CQAs (Arora, Gangu, & Jones, 2015; Hsu, Khbiri, & Caverlee, 2009; Ponzanelli, Mocci, Bacchelli, Lanza, & Fullerton, 2014b; Yao et al., 2015). Users of CQAs usually utilize these indicators to recognize high quality contents (i.e. Q&A) and high quality experts. As an example, intuitively, the probability of reading an answer is enhanced by the number of its votes.

Detecting and promoting high-quality contents and users is a primary success factor of a CQA which can lead to its overall quality and accordingly improves the site reputation, user experience, user engagement, and boost web result rankings.

Finding high quality contents and users (i.e. experts) on CQA is a well studied problem. Several aspects of this problem are investigated which can be divided in three main categories:

1. **Feature identification and selection**: The main approach in this category is to use a classifier to classify a given content (i.e. question, answer) or user into high or low quality item (Agichtein, Castillo, Donato, Gionis, & Mishne, 2008; Blooma, Goh, & Chua, 2012; Molino, Aiello, & Lops, 2016; Ponzanelli et al., 2014b; Toba, Ming, Adriani, & Chua, 2014; Yao et al., 2015). Here, the main goal is the identification and selection of important discriminative features. Several features like textual (Agichtein et al., 2008), temporal (Figueroa, Gómez-Pantoja, & Herrera, 2016), semantical (Figueroa & Neumann, 2016) and behavioral (Fu et al., 2016) have been proposed to detect the high quality content and users in CQAs. The methods in this category use separate classifiers for questions, answers and users. In other words, the quality label of each item (i.e. question, answer and user) is determined independent of other items.

2. **Quality dependency modeling**: Questions, answers and users are the main components of a CQA. The approaches in this line of research (Bian, Liu, Zhou, Agichtein, & Zha, 2009; Yao et al., 2015) utilize the quality dependencies between these components to improve the accuracy of the high quality content detection. For example, (Bian et al., 2009) uses the circular quality dependency between user, question and answer to improve the accuracy of prediction. This research assumes that an expert user usually provide a high quality answer in response to a good (i.e. high quality) question which is asked by an experienced user. Similarly, Yao et al. (2015) recognized a strong correlation between the number of votes on a question and its best answer and used it to simultaneously detect high quality questions and answers on CQA.

3. **Early detection of High quality items**: More recently, a new variation of high quality content detection problem on CQA is emerged in which the main goal is to predict of high quality posts (or users) in a short time soon after their submission (or user activity). For example, van Dijk, Tsagkias, and de Rijke (2015); Pal, Farzan, Konstan, and Kraut (2011) proposed a method to detect expert users in Stack Overflow after few weeks of their activity. Similarly, the problem of early high quality post detection has been recently introduced in Yao et al. (2015). Here, the main idea of research is to recognize and utilize time invariant patterns which can help to identify high quality items soon after their submission.

The main idea proposed in this paper can be categorized in number two and three of the above mentioned categories. Specifically, by investigating large amount of CQA data, we recognize two simple and important quality dependency patterns which can be used to enhance the high quality post detection problem. These two patterns are time invariant\textsuperscript{5} and we show that they can be used to early detection of high quality posts on CQA.

We called our first observed pattern on CQA data as the Accepted Answer Effect. According to this pattern, the quality of a question (i.e. its number of votes) is strongly dependent on whether that question has (or will get) an accepted answer or not. Intuitively, questions with accepted answer attract more attention from the community in comparison with other questions, because a question with accepted answer is usually relevant, important and general enough to motivate someone to answer it.

According to the result of our experiments, we show that this pattern is also valid in a short time period after submission of a question on CQA and accordingly, we use it for early high quality post detection.

In addition, while in previous research (Yao et al., 2015), two classifiers\textsuperscript{6} (one classifier to detect high quality questions and another one for high quality answers) have been used to detect high quality questions and answers, our contribution in this part is to utilize another classifier to predict whether the question will get an accepted answer in a short time after

\textsuperscript{1} answers.yahoo.com
\textsuperscript{2} www.quora.com
\textsuperscript{3} www.stackoverflow.com
\textsuperscript{4} math.stackexchange.com
\textsuperscript{5} By time invariant, we mean that these patterns are valid soon after the submission of a post on CQA.
\textsuperscript{6} These two classifiers can be separate or relational.
دریافت فوری
متن کامل مقاله

امکان دانلود نسخه تمام متن مقالات انگلیسی
امکان دانلود نسخه ترجمه شده مقالات
پذیرش سفارش ترجمه تخصصی
امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
امکان دانلود رایگان ۲ صفحه اول هر مقاله
امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
دانلود فوری مقاله پس از پرداخت آنلاین
پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات