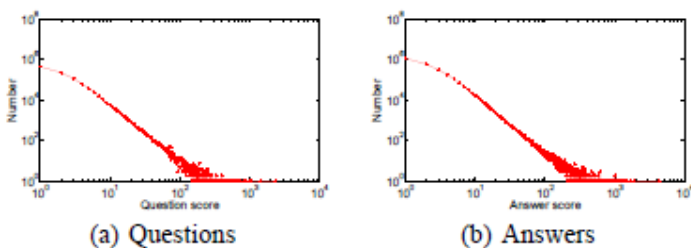




Data Mining Reveals the Secret to Getting Good Answers

If you want a good answer, ask a decent question. That's the startling conclusion to a study of online Q&As.



If you spend any time programming, you'll probably have come across the question and answer site [Stack Overflow](#). The site allows anybody to post a question related to programing and receive answers from the community.

And it has been hugely successful. According to Alexa, the site is the 3rd most popular Q&A site in the world and 79th most popular website overall.

But this success has naturally led to a problem—the sheer number of questions and answers the site has to deal with. To help filter this information, users can rank both the questions and the answers, gaining a reputation for themselves as they contribute.

Nevertheless, Stack Overflow still struggles to weed out off topic and irrelevant questions and answers. This requires considerable input from experienced moderators. So an interesting question is whether it is possible to automate the process of weeding out the less useful question and answers as they are posted.

Today we get an answer of sorts thanks to the work of Yuan Yao at the State Key Laboratory for Novel Software Technology in China and a team of buddies who say they've developed an algorithm that does the job.

And they say their work reveals an interesting insight: if you want good answers, ask a decent question. That may sound like a truism, but these guys point out that there has been no evidence to support this insight, until now.

“To the best of our knowledge, we are the first to quantitatively validate the correlation between the question quality and its associated answer quality,” say Yuan and co.

These guys began their work by studying the entire corpus of questions and answers on Stack Overflow between July 2008 and August 2011. That’s some 2 million questions from 800,000 people who produced over 4 million answers and 7 million comments. They also considered metadata, such as the number of upvotes and down votes for each entry.

Until now, most attempts to evaluate the quality of user input have looked only at the votes associated with questions or the votes associated with answers. For example, a good answer has more upvotes than downvotes and the bigger the difference, the better the result.

But Yuan and co dugged a little deeper. They looked at the correlation between well received questions and answers. And they discovered that these are strongly correlated.

A number of factors turn out to be important. These include the reputation of the person asking the question or answering it, the number of previous questions or answers they have posted, the popularity of their input in the recent past along with measurements like the length of the question and its title.

Put all this into a number cruncher and the system is able to predict the quality score of the question and its expected answers. That allows it to find the best questions and answers and indirectly the worst ones.

There are limitations to this approach, of course. First, it can only make its prediction after the first 24 hours of responses to a question. That’s not so useful to Stack Overflow since it needs to find ways of filtering out the lower quality questions before they reach the broader community. So Yaun and co say they are working ways to filter out the worst questions more quickly.

Second, Yuan and co rely on the impressive amount of metadata that Stack overflow collects for both questions and answers. That’s in stark contrast to many Q&A sites that allow users only to vote on answers. The moral for these sites may be to collect more data on the questions.

In the meantime, users of Q&A sites can learn a significant lesson from this work. If you want good answers, first formulate a good question. That’s something that can take time and experience.

Perhaps the most interesting immediate application of this new work might be as a teaching tool to help with this learning process and to boost the quality of questions and answers in general.

Ref: <http://arxiv.org/abs/1311.6876>: Want a Good Answer? Ask a Good Question First!

Tagged: Computing, Web

Reprints and Permissions | Send feedback to the editor
