Predicting Content Quality in Community Question Answering

Martin BORÁK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
xborakm@stuba.sk

Information age makes it easy for people to seek information on various topics. Everyone who has access to the Internet can just type anything that he is curious about into a search engine, and will receive thousands of results. However, sometimes people have more complex questions or they seek very specific answers, which are hard to find using traditional methods. Community Question Answering (CQA) systems are systems, where users can publicly ask questions and let other users answer them.

Content in CQA systems is user-generated, and therefore the quality of this content (questions, answers) is not always optimal. Consequently, a content quality prediction and evaluation process is necessary. Users need to be able to distinguish between content of high quality and that of low quality. It gives extra feedback, it enables them to solve their problems more quickly and comfortably and it also helps them produce better content in the future.

There are several papers, which try to solve the problem of content quality prediction and evaluation. Authors in [1] evaluated answer quality using graded relevance based on information available after some time since the answer was posted (community feedback). Research in [2] tried to predict which answers will be voted best using logistic regression. Methods developed in these papers give us better knowledge of what information effects answer quality the most and how can it be measured, but their helpfulness to everyday users is rather limited.

In our project, we specifically focus on answer quality prediction. That is, using features which are immediately available after an answer is submitted (text features, information from user profiles) we devise the answer's quality as the predicted number of votes that the answer will eventually get.

Predicting answers quality immediately after its creation is helpful to the asker, who can have a better perspective on which answer to trust, and also to the answerer, who, in case the quality of his answer will be determined to be low, is be able to edit or delete it.

---

* Supervisor: Ivan Srba, Institute of Informatics and Software Engineering

Spring 2015 PeWe Workshop, April 11, 2015, pp. 5–6.

Working with dataset from Stack Exchange and Askalot infrastructure, we have plenty of data and an excellent environment to test our method appropriately.

First thing needed in our method is feature extraction. We can work with text features, such as title length, main text length, usage of punctuation, referencing, usage of code snippets and so on. Also we are able to gather information from users profile, for instance how many answers did the user provide in the past, how many of them were marked as best, how long has he been registered in the system etc.

Using linear regression, which is a form of supervised machine learning, we then calculate weights of importance of these extracted features and determine the result, which will be a number representing prediction of future count of votes. We compare this result to the actual vote count that can be extracted from the dataset and in this way we are be able to evaluate the success rate of our method.

Despite the fact, that our main goal is to predict answer quality, the by-product of our method (weights of features) is also useful. Knowing what affects the quality not only helps us predict it with higher precision, but also enables us to recommend more concrete ways to improve answers with low quality.

We believe that our method has a potential to improve the user experience in CQA systems for users in ways of recognizing high quality content and also helping in producing it.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. 009STU-4/2014.

References

[1] Sakai, T. et al.: *Using graded-relevance metrics for evaluating community QA answer selection.* In Proceedings of the fourth ACM international conference on Web search and data mining - WSDM ’11. New York, New York, USA: ACM Press, (2011), pp. 187–196.

[2] Shah, C. & Pomerantz, J.: *Evaluating and predicting answer quality in community QA.* In Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’10. New York, New York, USA: ACM Press, (2010), pp. 411–418.