Recommendation of Solved Questions from Archives in CQA Systems

Viktória LOVASOVÁ*

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

Community Question Answering (CQA) sites such as Yahoo! Answers or Stack Overflow have become valuable platforms to create, share and seek a massive volume of human knowledge. To prevent information overload of users, we propose a method for personalized recommendation of already solved questions by personalized prediction of questions' information value which is usually expressed by favouring these questions.

The primary role of CQA systems is to answer new questions. In general, there are two main sources of knowledge for answering new questions:

- Community – CQA systems can recommend expert users who would know the answer on a new question for the user who asked the question, so he/she gets the best possible response.
- Archives of solved questions – CQA systems can search for similar questions that have been already solved and thus the user does not have to wait for the answer.

CQA systems such as Stack Overflow and Quora have users who have an extensive knowledge in the domain they participate in. One of the implications is that these systems contain content that has long-term value (information value) and can be recommended to people looking for answers to such questions. Authors in [1] worked on predicting long-term value of the questions and found out that the typical question answering consists of fast and slow phase. In the fast phase, it gains answers and votes and in the slow phase, members of the community indicate the question's long-term value by two events: visiting the question and by the mechanism of favouring the question. In the experiment described in [1], a number of question views with its answers served as an indicator of long-term value.

Predicting the question's long term value was also investigated in [3]. Authors designed a family of algorithms to solve prediction problem by modelling three key aspects: non-linearity, question/answer coupling and dynamics. They chose for their

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

Spring 2016 PeWe Workshop, April 2, 2016, pp. 41–42.

work user score because they did a survey in which they asked various users what they consider as the best indicator of long-term value in a CQA system.

In both studies [1] and [3], the authors considered the votes, favorites and question views as expression of interest from users. We took favorites in this analysis to see when users favor a question and to validate that a question contains a high information value for a particular user who favored it. Our goal is to propose a method which provides a user with recommendation of questions which have information value for him/her what he/she usually expresses by favorites. We decided for an offline evaluation on the data from CQA Android Stack Exchange because the authors from the earlier works [1] and [2] used frequently the data from the Stack Overflow to validate their methods. The first prototype of the method (with limited number of input features) was verified in two phases:

1. Problem of classification. In the first phase, we predict whether a user would favor a particular question or not. The questions were divided into two groups and we predicted class to which the questions belong to. We used precision, recall and accuracy as metrics. In precision we achieved 1.0, recall 0.97833 and accuracy 0. 9891. The obtained results indicate high successfulness of prediction, however, it is partially the result of the extreme cases (user favors or votes for close) that we have classified and their different values of features.
2. Learning to rank problem. From the first phase we evaluated that we can identify the favorite questions with a high precision so in the second phase, we simulated a real usage of our proposed method in the CQA system. A typical scenario for recommendation would be that the users would every week get recommendations with top 10 questions that have information value for them. The recommendations would be calculated correctly, if the list of recommended questions contain the question which the user really favored. In this scenario, the results did not achieve so high performance, we managed to generate 74% correct newsletters (question which the user favored was present in the list).

Extended version was published in Proc. of the 12th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2016), STU Bratislava.

References

[1] Anderson, A., Huttenlocher, D., & Kleinberg, J.: Discovering Value from Community Activity on Focused Question Answering Sites : A Case Study of Stack Overflow. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2012), pp. 850–858.

[2] Ji, Z., & Wang, B.: Learning to rank for question routing in community question answering. Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management - CIKM '13, (2013), pp. 2363–2368.

[3] Xu, F., & Lu, J.: *Predicting Long-Term Impact of CQA Posts : A Comprehensive Viewpoint,* (2014), pp. 1496–1505.