Full length article

A hybrid model to predict best answers in question answering communities

Dalia Elalfy *, Walaa Gad, Rasha Ismail

Information Systems Department, Faculty of Computer & Information Sciences, Ain Shams University, Abbasia, Cairo, Egypt

ABSTRACT

Question answering communities (QAC) are nowadays becoming widely used due to the huge facilities and flow of information that it provides. These communities target is to share and exchange knowledge between users. Through asking and answering questions under large number of categories.

Unfortunately there are a lot of issues existing that made knowledge process difficult. One of those issues is that not every asker has the knowledge and ability to select the best answer for his question, or even selecting the best answer based on subjective matters. Our analysis in this paper is conducted on stack overflow community. We proposed a hybrid model for predicting the best answer. The proposed model is consisting of two modules. The first module is the content feature which consists of three types of features; question-answer features, answer content features, and answer-answer features. In the second module we examine the use of non-content feature in predicting best answers by using novel reputation score function. Then we merge both of content and non-content features and use them in prediction. We conducted experiments to train three different classifiers using our new added features. The prediction accuracy is very promising.

© 2017 Production and hosting by Elsevier B.V. on behalf of Faculty of Computers and Information, Cairo University. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

There are many types of social networks that you can use through internet. The types are according to the functionality each network can provides [1]. One of these social networks is collaborative social network. Stack Overflow (http://stackoverflow.com), Quora, Yahoo!-answers and Never are examples of this type. In stack Overflow community the question answering process is conducting as follows. The user can choose a category to post a question to. After posting the question the asker waits specific amount of time in order to receive the answers from the expert users. Expert users are the users whose have a great knowledge in this category or sub category of the question field. If the question does not receive any answer, the asker can set a bounty to it. A bounty is a special reputation award given to answers. This feature was designed to motivate answerers, and help questions get the answers they deserve. Bounty awards are funded by the personal reputation of the users who offer them. Reputation is a rough measurement of how much the community trusts the answer author; it is earned if the answer author convinced other users that his answer is the best and this answerer knows what he is talking about. While bounty maker do not need to be the owner of a question to start a bounty on that question, only one bounty can be active on a question at once, and each user can only have up to three active bounties at once. Users must have specific reputation score to start a bounty, and at least as much reputation as the bounty amount. The bounty award will be subtracted from your reputation when the bounty is started, not when it is awarded.

If the question receives many answers, the users can give an up and down votes to both the question and answers in its answer thread. The most reputation points score is gained when the answer is up voted, it received a bounty, or it is selected as best answer. Also users can add comments to question and answers. Moreover users can set post as a favorite post. There are other activities but to be able to use them it depends on a user privilege under the community. User privilege depends on user reputation score. As an example of these privilege is to mark question or an answer as an offensive post.

* Corresponding author.
  E-mail address: dallol_elalfy47@hotmail.com (D. Elalfy).

Question's author can then select the most preferable or satisfied answer. This answer is called the best or the accepted answer. Unfortunately, this mechanism in question answering portals may lead to a lot of issues. Such as some askers cannot be able to choose the accepted answer.

For their question. The consequences are a lot of questions left as "not-answered" [2]. Even if the question is answered, there exist some probability that the answerer is not an expert in such category. And just answer it because he faced with that question when he/she opens the question answering portal. As a result the community lose one of its targets which is sharing knowledge because of the low quality answers that can be exists.

Moreover, the answerer that gives the low quality answer is capable of give a high quality one if that user was faced by the right question that he/she is expert in.

In this paper, we focus on the problem of exchanging and sharing of the knowledge in stack overflow community. And how to ease knowledge exchange by saving asker wasted time and effort that user exerts to find a satisfied answer. By predicting the best answer to the user, which considered the main problem in question answering portals.

The work in this paper is organized as follows. In Section 2, related work and a depth analysis to different knowledge exchange approaches will be introduced. Section 3, a study for the model in Ref. [3] which is used to give a local reputation score to answers and our proposed hybrid model for predicting best answer. Section 4, presents the experiment results and discussion. In Section 5 the conclusion is presented.

2. Related work

In order to improve the mechanism in which question answer portals work under, we need to focus on enhancing the question answer routing approaches and finding best answer techniques... There are a lot of efforts done by researchers in this field to overcome these problems; we are categorizing these studies to four categories: recommend right experts to a specific question, predicting the best answer, finding group of collaborative experts, and direct questions to an expert. All of these are solutions in order to improve the user satisfaction rate by giving high quality answers to him. Also to minimize the loss of time that is as a result of waiting for the right expert to answer the question.

2.1. Recommend experts to the current question

In Ref. [4] they tend to find the right expert to answer specific question under certain category. They proposed a hybrid model to find experts using user reputation, user authority, and user subject relevance. In evaluating their model they used Yahoo! Answer platform in Taiwan. Also called Yahoo! Knowledge plus. They assign different priority to terms according to their place like if the word is in answer post, question post, or in question title. One of the main issues in their technique that they do not consider the quality of posts posted by the expert. Since HITS take only the number of posts as an indication to the authority of that user.

In Ref. [5] their aim is to recommend an appropriate users to answer a specific type of. They split questions into two types an authority and affinity questions. They also recommend

A social network that is suitable to answer that type of question. Authors created a website that would allow user's from different social networks to ask questions in order to gain knowledge or social interaction. Social networks used are Facebook, Myspace, and Twitter. The authority question is the question that seeks information, the affinity question is the question that seeks social interaction or opinion. They build Expertise Estimation Algorithm

to determine the objectivity levels of questioners and responders. This objectivity level is used later in determine the affinity and authority users. They does not consider the relationship between answer and question so that the answer might be off-topic.

In Ref. [6] Enterprise Social Network (ESN) service can help employees to collaborate and communicate effectively with colleagues, with customers and with suppliers. In this paper authors propose a model to better support question answering process in ESN, by using a graph analysis approach. Based on the questioner's initial input list of potential answerers, it can extract a shared-interest group of people, whose interest is similar to the initial list of potential answerers, and sort the group of people according to a score of interest distance, and then recommend them to the questioner. To evaluate its applicability, the method is implemented in KDWeibo the most popular ESN platform in China. The algorithms include three key concepts: Interest Distance, Aggregate Specialization Graph (ASG) and Specialization Sub Graph (SSG). One of the drawbacks exist in their model is that the method will not work if the user is not providing any initial list of potential answerers or the quality of the initial list is too low.

The Authors in Ref. [7] introduced a probabilistic framework to predict best answerers for questions. By tracking answerers' history, interests of answerers are modeled with the mixture of the language model and the Latent Dirichlet Allocation Model LDA. They also used both user authority and activity in predication. They utilizes two models to define user's interests. They calculated the likelihood probability based on user profile in order to predict the expert. Also they model the prior information of user which is made up of two parts the authority and user activity. They conducted their experiment using Isak CAQ in china. As a feedback to their work they need to further investigate their work in large scale dataset. And to use more accurate and different user activity and authority models.

In Ref. [8] Authors use a mechanism to filter online social streams received as well as enable them to interact with most similar users by personalizing the process of information distribution. Their framework identifies the most appropriate users to receive specific post by calculating the similarity between posts of the target user and the others. The platform used in research is stack overflow. Similarity is calculated based on user's social activity. User's social activity is an integration of both user interest in posts published and social activities of that user. Each user is represented by two vectors in vector space model. First vector is social pattern vector which contains influence attributes and user's distribution. The second vector contains bag of words as a post content's vector. Term frequency and inverse document frequency is used to weight terms of each vector. Then an aggregated linear model is applied to combine the calculated cosine similarity in two vectors.

2.2. Finding the best answer

In Ref. [9] they focus on finding best answer in massive online open courses in which users enroll in courses and to further understand it they can ask and answer question in the course forum. The experiment is conducted on openHPI MOOC platform. The users used machine learning through train four classifiers. They are bagging, naive Bayes, MultiPerceptron, and Random Forest using user features, thread features and content features. They used as a historical data the questions that has at least two answers. The training is performed on the answers of 416 questions.

Ref. [10] is a survey that the researchers found that there is a high correlation between posting a high quality question and getting a high quality answer. So they studied the features that most important in question to be found in order to get high quality answers. These question related features are tags and terms, length of the question, presence of an example that may help users to