Stack Overflow: Mining domains with lack of answers, and building a recommendation system

•••

Fall 2015

Project No : 7

Abstract

Our research has found out that there are many domains on stack overflow where the answer seekers are many(there a lot of questions asked in that domain) but the answer givers are less. We found those domains and then found the users with similar answering patterns. Other domains will lead to the domains of interest and thus the recommendation.

This means that domains, for which a large number of questions were asked and little number of answer givers, will be recommended and hence people answer them and this can be incorporated in stack overflow website by giving some additional credits. This way all the domains will start getting answered and stack overflow will become even more successful.

Understanding the problem

No answer givers

There are many domains in which the answer seekers are more and answer givers are less.

User's answering pattern

The user's answering pattern are the most crucial point in building the recommendation system.

Attracting Users

We need to attract users to give answers in the domain which have lack of answers, so that there is an increase in their reputation.

Project objective:

To find the domains where the ratio of answer giver to answer seeker is quite low and then find the frequent item set where these domains fit it and then recommend the other domain pattern in the set these domains and thus get them answered. People should be recommended the domains that are “hot” and requires people to give answers.

Steps in our algorithm

Step 1

Finding the domain patterns where there were a lot of questions. On the question data set we find the tag patterns and then apply fp-growth algorithm for frequent tag patterns count. This dataset consisted of over 4 lac questions asked in a period of 1st Jan,15 to 1st May,15.We got the list of all the domains in which the number of questions asked was more than 300

Step 2

We find the count to these tag patterns in the questions and then find the count in the answer data set and then find the ratio = |is_answered|/|questions with the pattern.We took the 20 tag patterns with least answering ratio. Excluded zero ratio ones as the ratio zero means that no answer was given so we cannot find the similar domains from the answer set.

Step 3

For each of the 20 domains, we found the frequent answerers, and also looked at their answering patterns across different domains. Then for all the frequent answerers in a domain we found out the frequent answering patterns, which helped us in finding out the domain to which we could recommend the domain in consideration.

Testing

To test our recommendation system we used the follwing method.

Consider a domain ‘d’. Let the set of all the domains to which questions of ‘d’ can be recommended be ‘D’. For each domain in ‘D’, we found out the frequent answerers in the period 1st May,15 to 1st Aug,15. We found out that 30 percent of these users are answering the questions of ‘d’, and also the percentage of users answering the questions in ‘d’, but not answering any of those in ‘D’ was 40 percent. Hence the percentage of questions answered by the users who are also answering the domains in ‘D’, and also the ones in ‘d’, is 60%. This recommendation system will help in increasing this 60% and also the total number of users answering questions of ‘d’.

Future

The type of recommendations in stack overflow can be divided into two types with two objectives:

1. Personalized recommendation :
   Finding similarity between users, questions and then recommending.
2. Domain cover recommendation :
   Finding similarity between users, questions and recommending domains which remain unanswered and lack expertise of answer givers. This can be incentivized and thus all the questions will be answered and more popularity of stack overflow.

THANK YOU!

...

Kartik Gupta
Naman Singhal
Vishal Gupta
Vishal Thamizharasan