Expert Mining in Stack Overflow

(Empirical Track)

Ankit Joshi (Unity ID – ajoshi5), Tanvi Mainkar (Unity ID – tmainka)

Introduction

Stack Overflow is a widely popular Q&A forum which is being used by programmers worldwide. It provides a platform to discuss and seek solutions to various problems in variety of technologies. Since StackOverflow has a very active user base, majority of the problems receive multiple solutions, and that too quickly. However, the most common concern, especially for new users on StackOverflow, is whether or not they can trust the solutions provided by other users. A new user can safely trust a solution only if it has been posted by someone, who can be thought of as an expert in the field. Thus, determining the expertise of a user in a specific technology is very important, especially from the point of view of a new user to StackOverflow. This project explores a method, which would help a user to determine whether another user is an expert in a certain technology. The remainder of this proposal will describe the motivation and the project details as well as mention the tools that would be used for the project.

Motivation

On StackOverflow, the reputation of another user serves as a great indicator of his level of expertise in different technologies. A StackOverflow user primarily gains reputation when other users find his answers useful and vote for it. Thus, a user earns majority of his reputation score by the number of votes he has gained by answering questions on a variety of technologies. However, reputation is not a clear indicator of whether a person is an expert in a specific technology. We can try to determine whether a user is an expert in a particular technology by the number of points he has gained by answering questions related to that technology.

Project Summary

This project, based on the analysis of StackOverflow's data, will help in determining whether a StackOverflow user, who has gained points above a certain threshold in a particular technology, can be thought of as an expert in that technology or not.

Project Details

The activity of a set of users would be collected from the StackOverflow data dump. Users are assumed to be experts in a particular technology, if they have earned points above a certain threshold in that technology. This data set will be analyzed to determine these users' average response times to the questions posted, the number of their answers which got accepted, the number of answers which received the highest points and the number of badges they have earned in their area of expertise. The data obtained from the above analysis will be used to determine whether the assumption made about the user being an expert in that field is correct or not.

Project Environment

This project will make use of the official data dump provided by Stack Overflow for 2014. (https://archive.org/details/stackexchange). A local copy of the StackOverflow database will be created in MySQL using their official data dump. The code for data analysis and comparison would be written in Java.

Conclusion

Reputation on StackOverflow by itself is not a clear indicator of whether a user is an expert in a particular field or not. We might be able to determine whether a user is an expert in a particular field, by the points he has gained in that field. Therefore, this project makes an assumption, that a user is an expert in the field, if he has earned points above a particular threshold in that field. The project will then try to prove this assumption based on certain parameters that usually determine expertise in a field. If the assumption is proved to be true, then a user can determine which other users on StackOverflow are experts in a particular technology. The askers/inquirers can then almost safely consider the expert users' answers to be true.

References

1) StackOverflow – What is Reputation?- http://stackoverflow.com/help/whats-reputation