Project Proposal: Analysing Stack Exchange Forums

Valeria Chernenko
xxx@in.tum.de

Kerstin Gottelt
xxx@in.tum.de

Oleksandr Shchur
oleksandr.shchur@in.tum.de

Tizian Sarre
tizian.sarre@gmx.de

1. DATA SET

The data set we are intending to work on contains information about user activity on Stack Exchange forums. Stack Exchange is a network of Q&A boards on topics such as Software Engineering, Linux Administration and Mathematics. The data set itself is divided into parts for each of the boards, each of the parts consisting of data about posts (a.k.a questions) tagged according to different topics, answers, user profiles, comments and votes.

2. FINDING DUPLICATE POSTS

Our first idea is to develop an algorithm to find highly similar posts. One possible application for it will be to check new posts for uniqueness. It happens often that inexperienced users ask questions that have already been answered. In this case pointing directly to the answer will save time for the user and will reduce the amount of redundant information on the forums. Currently, the system only checks for similarity based on the title of the question, which can most definitely be improved upon by also taking into consideration the body and the tags.

Another potential way to apply our solution will be to suggest tags for new posts that only have few or none at all. Doing this will make it easier for experts to find the questions they might have the knowledge to answer. This will allow the askers to get faster responses, make navigation easier for experienced users and, hopefully, increase the rate at which questions are being answered. As of now, this feature is not implemented on Stack Exchange.

3. MILESTONES

To keep track of our work progress, we'd like to set following milestones:

1. Design a smart similarity measure that works well in the context of Stack Exchange posts.
2. Implement an algorithm that outputs a list of similar posts for any query post.
3. Apply the results from (2) to find duplicate posts.
4. Implement the tag suggestion system for posts that have none or few.