Project Proposal: Analysing Stack Exchange Forums

Valeria Chernenko
valeriia.chernenko@tum.de

Kerstin Gottelt
ke.gottelt@in.tum.de

Oleksandr Shchur
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. MOTIVATION

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. The current realisation lists possible duplicates based only on the title of the question. Our approach will take into account not only the title, but also question's body and tags. It will most likely improve the quality of analogous questions' detection and help users to avoid posting duplicates.

Second possible application of our technique is to provide links to similar questions for existing ones that are not yet checked as answered. Sometimes the questions could not be resloved by other users due to a lack of detaills or bad question formulation. If there are other similar questions that already have a solution, pointing a link to them could help finding correct answer for original question. That will also allow the users who reach StackExchange via a search engine to find the solutions faster.

Another potential way to apply our solution will be to suggest tags for new posts that only have few or none at all. For instance, on StackOverflow currently more than 5 out of 9 million posts have no tags assigned to them. Resolving this issue 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, the StackExchange platform suggests tags only if there are hyperlinks in the question's body. A question with no links included doesn't receive any tag recommendations at all.

3. MILESTONES

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

1. Design a similarity measure that works well in the context of Stack Exchange posts.
2. Implement an algorithm that outputs a list of similar posts and corresponding similarity measure for any query post.
3. Design a metric to compare our results to current StackExchange ones.
4. Apply the developed algorithm to find duplicate posts on StackExchange forums of different sizes. Compare the results.
5. *(optional)* Implement the tag suggestion system for posts that have none or few.