Data Exchange and Preparation for Mining StackOverflow and GitHub Data

COMP 5900 Project Proposal

Norbert Eke
Carleton University
Ottawa, ON
NorbertEke@cmail.carleton.ca

KEYWORDS

Data Mining, Mining Software Repositories, Stack Overflow, GitHub, Data Cleaning, Data Mapping, Topic Modeling

ACM Reference Format:
Norbert Eke. . Data Exchange and Preparation for Mining StackOverflow and GitHub Data: COMP 5900 Project Proposal. In *Proceedings of* . ACM, New York, NY, USA, 2 pages.

1 INTRODUCTION

StackOverflow is the most popular question answering website for software related questions. In the latest public data dump from December 9th 2018 StackOverflow listed over 42 million posts from almost 10 million registered users. To analyze how StackOverflow posts evolve Baltes et al. [2] built SOTorrent [1], an open data set aggregating and connecting the official StackOverflow data dump to other web sources, such as GitHub repositories. SOTorrent provides access to the version history of StackOverflow content at two different levels: whole posts and individual text or code blocks. GHTorrent [3] is similar to SOTorrent, as the creators wanted a queriable offline mirror for the data present on Github.

2 PROBLEM STATEMENT

The problem of developer expertise recommendation is a well defined problem in software engineering. Project managers are often faced with the task of deciding who is the best developer to handle a certain bug fix, code review or pull request. A task at hand needs to be assigned to one or more developers in the company. The optimal task assignment would allow the developer(s) with most expertise and experience in the domain to complete the task at hand. To achieve such a task assignment, predicting the expertise of a developer is needed.

3 MOTIVATION

Topic models are a type of statistical model used in text mining to discover hidden semantic structures in a textual data. Most of such models discover patterns of distributions of topics within the textual data. Tian et al. [4] used topic models in their work on predicting the best answerer for a new question on StackOverflow. Their approach learns user topical expertise and interest levels by profiling each user's previous activity and reputation on StackOverflow. Tian et al. [4] claim that the "semantic similarity between the user profile and the new question can be captured through topic model". For each potential answerer each user's expertise level can be learnt through previous user activity data and up votes from the user profile. What my research is interested in find is developer (user) expertise level on StackOverflow, and potentially Github combined. The Naive approach of counting upvotes on each user's answers and associating them to tags on the questions would not work. Most experts in the community of software repository mining would agree that tags on StackOverflow are too general to define expertise level. For instance, the most frequently used tag on StackOverflow is JS (JavaScript), but saying that some is in expert in JavaScript does not reveal what exactly they know within that language.

4 OBJECTIVES

The end goal of my research is to build a predictor model for developer expertise on GitHub (GH) and StackOverflow (SO) linked together (more about the linking later).

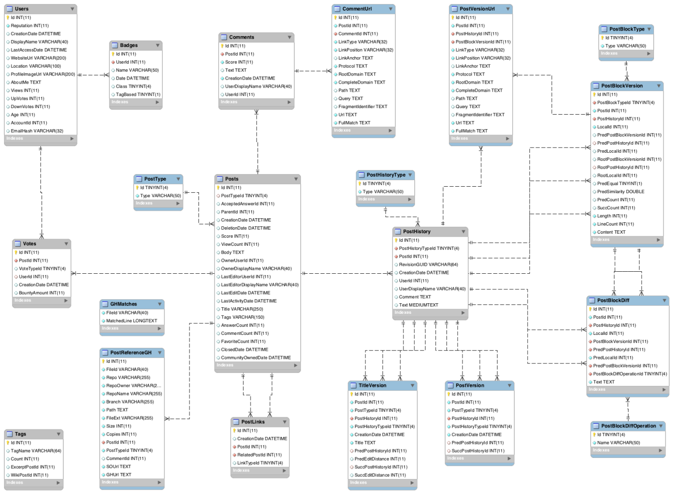

Figure 1: Database schema for the SO data set.

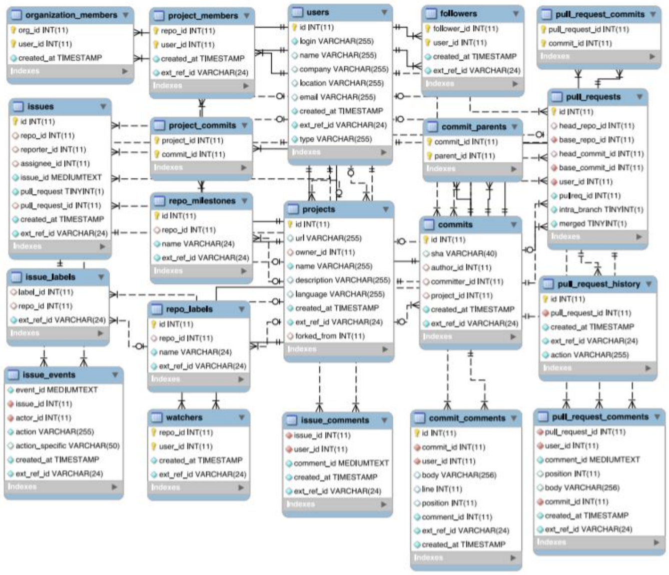

Figure 2: Database schema for the GH data set.

My research considers topic modeling to be essential for building such a predictor model, but data preparation is the largest task. Firstly, in order to link the GH and SO data sets, identity matching is required. Identity matching is the processing of matching usernames or aliases from one data set to another. This can be done for the users present in both GH and SO data sets by matching the MD5 hash of their email addresses provided in both data sets.

Data preparation will consist of 4 main tasks: identity matching, designing a new database schema, exchanging data between the source (GH + SO) and target (new) schema, then processing through all fields, and cleaning the data. The new database schema will have to contain only the desired relations from both data sets. Cleaning the data would consist of either changing or keeping the format of the non-textual fields (i.e. counting frequencies or aggregating numerical values) while thoroughly conducting text pre-processing on the textual fields.

In terms of data that will be useful for building the developer expertise model, user activity (answers to questions), user reputation (up votes) and user profile data (about me), as well as other data should be used.

Looking at the database schema of the SO data set in Figure 1, the relations named Comment, Post, User, Votes and Tags will be useful towards analyzing developer expertise levels.

Looking at the database schema of the GH data set in Figure 2, the relations named users, issue_comments, commit_comments, pull_request_comments, repo_labels and projects will be useful towards analyzing developer expertise levels.

Linking and mapping together GH and SO expertise level data is a much bigger contribution to the software repository mining community than just looking at GH or SO data on it's own. Comparing expertise levels and behaviours of developers across both platforms will create interesting research questions.

For this course project only the data exchange and data preparation stages are intended to be completed. If time allows it, some topic modeling algorithms could be executed to gain potential insight into topical distributions of the data.

REFERENCES

[1] Baltes, S., Dumani, L., Treude, C., and Diehl, S. Sotorrent: reconstructing and analyzing the evolution of stack overflow posts. In *Proceedings of the 15th International Conference on Mining Software Repositories, MSR 2018, Gothenburg, Sweden, May 28-29, 2018* (2018), pp. 319–330.

[2] Baltes, S., Treude, C., and Diehl, S. Sotorrent: Studying the origin, evolution, and usage of stack overflow code snippets. *arXiv preprint arXiv:1809.02814* (2018).

[3] Gousios, G. The ghtorrent dataset and tool suite. In *Proceedings of the 10th Working Conference on Mining Software Repositories* (Piscataway, NJ, USA, 2013), MSR '13, IEEE Press, pp. 233–236.

[4] Tian, Y., Kochhar, P. S., Lim, E.-P., Zhu, F., and Lo, D. Predicting best answerers for new questions: An approach leveraging topic modeling and collaborative voting. In *International Conference on Social Informatics* (2013), Springer, pp. 55–68.