Project Abstract

Group Members

Fidelia Orji, fao583, fao583@mail.usask.ca
Amirabbas Jalali, amj301, amj301@mail.usask.ca
Seyedeh Mina Mousavifar, sem311, sem311@mail.usask.ca, will hand in the abstract

Project Topic Title

Benching Machine Learning Methods for Expert Recommendation
Short title: Expert Recommendation

Project Short Abstract

The community question and answer (CQA) platforms, such as Stack Overflow, leverage the knowledge and expertise of users to answer questions posted by fellow users. Over time, these websites turn into repositories of knowledge. Knowledge acquisition and exchange are generally crucial yet costly for both businesses and individuals, especially technical knowledge which covers a wide range of topics. CQA platforms offer an opportunity for sharing knowledge at a low cost, where community users, many of whom are domain experts, can potentially provide high-quality solutions to a given problem. However, in the era of information explosion, a large volume of questions are posted every day, and it is challenging to identify the relevant qualified experts to answer these questions. In this project, we aim to recommend the experts who are most likely going to answer a given question in the Stack Overflow community. We will explore and benchmark various machine learning methods to perform expert recommendations. Specifically, we will first collect the Stack Overflow dataset via publicly available data archive. The dataset captures user information such as a user's past answered questions, the number of badges awarded, etc. Data on the questions are also captured in this dataset, e.g., the question text, questioner, answerers, etc. After pre-processing the collected data, we will construct user and question features, which will be used as input into various machine learning models for expert recommendations. We aim to benchmark machine learning methods such as Naive Bayes, Support Vector Machine, Logistic Regression, Decision Tree, and Random Forest for the recommendation. We will evaluate the performance of the models using commonly used ranking-based metrics such as NDCG and HitRate. Finally, we will also empirically analyze the importance of our engineered features on different methods.