Wisdom in Sum of Parts: Multi-Platform Activity Prediction in Social Collaborative Sites

Roy Ka-Wei Lee
Living Analytics Research Centre
Singapore Management University
roylee.2013@smu.edu.sg

David Lo
Singapore Management University
davidlo@smu.edu.sg

ABSTRACT

In this paper, we proposed a novel framework which uses user interests inferred from activities (a.k.a., *activity interests*) in multiple social collaborative platforms to predict users' platform activities. Included in the framework are two prediction approaches: (i) *direct platform activity prediction*, which predicts a user's activities in a platform using his or her activity interests from the same platform (e.g., predict if a user answers a given Stack Overflow question using the user's interests inferred from his or her prior *answer* and *favorite* activities in Stack Overflow), and (ii) *cross platform activity prediction*, which predicts a user's activities in a platform using his or her activity interests from another platform (e.g., predict if a user answers a given Stack Overflow question using the user's interests inferred from his or her *fork* and *watch* activities in GitHub). To evaluate our proposed method, we conduct prediction experiments on two social collaborative platforms widely used in software development: GitHub and Stack Overflow. Our experiments show that combining both *direct* and *cross* platform activity prediction approaches yield the best accuracies for predicting user activities in GitHub (AUC=0.75) and Stack Overflow (AUC=0.89).

ACM Reference Format:
Roy Ka-Wei Lee and David Lo. 2018. Wisdom in Sum of Parts: Multi-Platform Activity Prediction in Social Collaborative Sites. In *Proceedings of ACM Conference on User Modelling, Adaptation and Personalization (UMAP'18).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Software developers are increasingly adopting social collaborative platforms for software development. *GitHub* and *Stack Overflow* are two of such popular platforms. GitHub is a collaborative software development platform that allows code sharing and version control. Stack Overflow is a technical question-and-answer community based website where users post and answer questions relating to software development. As these social collaborative platforms gain popularity, many research studies have proposed recommender systems to improve the usability of these platforms. For example, there are works which predict and recommend relevant Stack Overflow questions and answers to aid users in software development [2, 11, 12, 14]. While for GitHub, researchers have proposed methods to predict which software repositories are more relevant to a target user [4, 5, 7, 13, 15, 16]. Nevertheless, many of these studies only consider the users' behaviors and interests in a single platform when predicting and recommending user platform activities.

UMAP'18, July 2018, Singapore
2018. ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

There are a number of benefits for using user interests from multi-platforms for activity prediction. Firstly, it enables prediction of user activities in social collaborative platforms even when past activity history of a user is minimal or unavailable, i.e, cold-start problem [8]. For example, if we learn from a user's activities in GitHub that she is interested in *Python* and *text mining* techniques, we would predict that she will likely participate in *Python* and *text mining* related Stack Overflow questions even when she has just newly joined Stack Overflow and has not participated in any questions. Second, it could cover the *blind spots* of activity recommender systems which use only data from a single platform. For example, if a user has forked *Android* related repositories in GitHub, recommender systems which are build on user's past activity in GitHub will likely to recommend the user more *Android* related repositories. However, the same user have also participated in some *iOS* related questions in Stack Overflow, and such observations can be used to make relevant GitHub activity recommendations to the user.

There have been few existing inter-platform studies on GitHub and Stack Overflow [1, 6, 9, 10]. In a recent study by Lee and Lo [6], the researchers found that users who have accounts on both GitHub and Stack Overflow do share similar interests across the two platforms. For example, a user who commits to Java repositories in GitHub, is likely to also answer Java questions in Stack Overflow.

In this paper, we extended the study in [6], and proposed a novel framework which enables predicting users' activities using interests inferred from their activities in multiple social collaborative platforms. Secondly, we evaluate our method using large real-world datasets from Stack Overflow and GitHub. The results from our prediction experiments show that our proposed method is able to predict users' activities in GitHub and Stack Overflow with good accuracy, achieving an AUC score of up to 0.75 and 0.89 respectively.

2 MULTI-PLATFORM ACTIVITY PREDICTION

Figure 1 shows the framework that we adopt for multi-platform activity prediction. We begin with data extraction from two social collaborative platforms: Stack Overflow and GitHub. There are three sub-processes in data extraction: (i) matching of users Stack Overflow and GitHub accounts, (ii) extracting the users' platform activities, and (iii) inferring users' interests from their activities. Next, we construct the Stack Overflow and GitHub user features which we will use in our prediction. Our framework also incorporates two approaches to predict users' platform activities, namely: *direct* and *cross* platform activity prediction. We define *direct platform activity prediction* as predicting a user's platform activity using features from the same platform. For example, we predict if a given user will answer a given Stack Overflow question using the user's

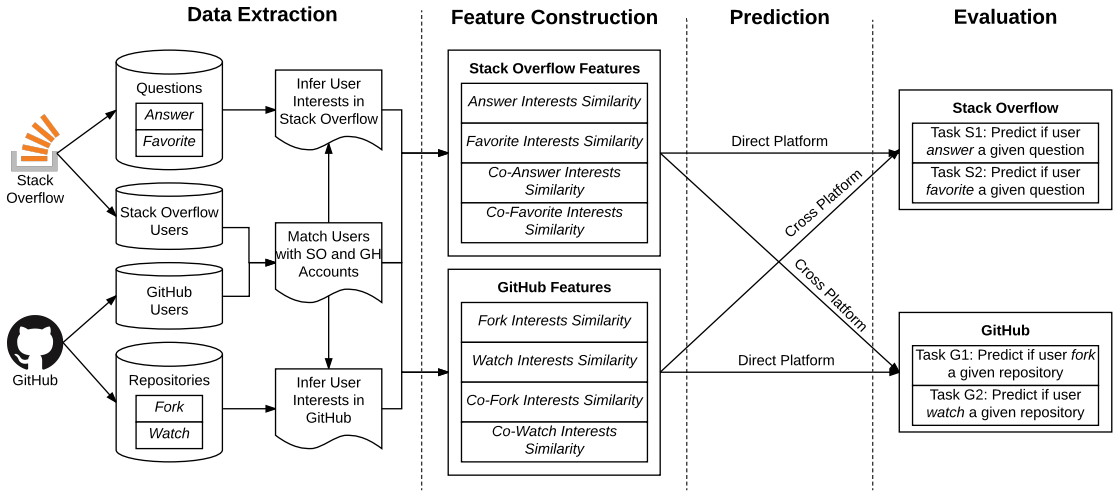

Figure 1: Multiple-Platform Activity Prediction Framework

Stack Overflow features. Conversely, we define *cross platform activity prediction* as predicting a platform activity to a user using features from a different platform. For example, we predict if a given user will answer a given Stack Overflow question using the user's GitHub features. The performance of both prediction approaches will be evaluated on four prediction tasks described in Section 4.

Table 1: List of notations used

Symbol	Description
u	Query user
k	Query item
v	User who co-participated activities with user u
r	Repository
q	Question
$I(r)$	Topics of repository r
$I(q)$	Topics of question q
$I(k)$	Topics of query item k
$u.RF$	Set of repositories forked by user u
$u.RW$	Set of repositories watched by user u
$u.QA$	Set of questions answered by user u
$u.QF$	Set of questions favorited by user u
$Co^{Fork}(u)$	Set of users who co-forked at least one repository with user u
$Co^{Watch}(u)$	Set of users who co-watched at least one repository with user u
$Co^{Ans}(u)$	Set of users who co-answered at least one question with user u
$Co^{Fav}(u)$	Set of users who co-favorited at least one question with user u

2.1 Problem Statement

Given a pair of query user and item (i.e., question or repository), (u, k), we aim to predict if u will perform an activity (e.g. answer, favorite, fork or watch) on k. There are various ways to measure the likelihood of u performing an activity on k. For example, we could consider the similarity between k's topics and u's topical interests inferred from different activities, or the similarity between k's topics and the inferred topical interests of user who co-participate activities with u. In our proposed framework, we propose two types of user features, namely:*user activity interest similarity features* and *user co-activity interest similarity features*. The notations used throughout this paper are summarized in Table 1.

2.2 User Activity Interest Similarity Features

This set of features measures the similarity between a query item k and a query user u's *fork*, *watch*, *answer* and *favorite* activity topical interests in GitHub and Stack Overflow. The intuition behind this set of features comes from the empirical study from Lee and Lo [6], where they found that users in GitHub and Stack Overflow shared similarities between their interests in different types of activities and across the two platforms. Suppose that we want to predict if a user would fork a given repositories in GitHub, we would measure the similarity between the given repository's topic and the user's topical interests for the different activity types. Intuitively, the higher the similarity scores, the more likely the user would fork the given repositories. Equation 1 captures the above intuition and measures similarity between k's topic and u's fork activity topical interests (i.e., $Sim_{Fork}(u,k)$), by dividing $\{r \in u.RF | I(r) \in I(k))\}$, which is the number of u's forked repositories that shared common topics with the item interests of k, by the total number of repositories forked by u (i.e.,$u.RF$).

$$Sim_{Fork}(u,k) = \frac{|\{r \in u.RF | I(r) \in I(k))\}|}{|u.RF|} \quad (1)$$

$$Sim_{Watch}(u,k) = \frac{|\{r \in u.RW | I(r) \in I(k)\}|}{|u.RW|} \quad (2)$$

$$Sim_{Ans}(u,k) = \frac{|\{q \in u.QA | I(q) \in I(k)\}|}{|u.QA|} \quad (3)$$

$$Sim_{Fav}(u,k) = \frac{|\{q \in u.QF | I(q) \in I(k)\}|}{|u.QF|} \quad (4)$$

We compute the similarities between k's topics and u's watch, answer and favorite activities topical interests in similar ways as shown in Equation 2, 3 and 4 respectively.

2.3 User Co-Activity Interest Similarity Features

This set of features measures the similarity between a query item k's topic and the activity topical interests of other users v who have co-participated in an activity with a query user u. The intuition behind this set of features also comes from the empirical study from Lee and Lo [6], where they found that users share similar interests with other users who they co-participated an activity (even minimally) in social collaborative platform. Suppose that we want to predict if a user would fork a given repository in GitHub, we would measure the similarity between the given repository's topic and the topical interests of other users who had co-forked repositories with the user in GitHub. Intuitively, we would also expect that the higher the similarity score, the more likely the user would answer the given question. Equation 5 captures the above intuition and measures the average similarity between k's topic and fork activity topical interests of all users v, who had co-forked at least one question with u (i.e., $Co^{Fork}(u)$).

$$Sim_{CoFork}(u,k) = \frac{\left[\sum_{v \in Co^{Fork}(u)} \frac{|\{r \in v.RF | I(r) \in I(k)\}|}{|v.RF|}\right]}{|Co^{Fork}(u)|} \quad (5)$$

$$Sim_{CoWatch}(u,k) = \frac{\left[\sum_{v \in Co^{Watch}(u)} \frac{|\{r \in v.RW | I(r) \in I(k)\}|}{|v.RW|}\right]}{|Co^{Watch}(u)|} \quad (6)$$

$$Sim_{CoAns}(u,k) = \frac{\left[\sum_{v \in Co^{Ans}(u)} \frac{|\{q \in v.QA | I(q) \in I(k)\}|}{|v.QA|}\right]}{|Co^{Ans}(u)|} \quad (7)$$

$$Sim_{CoFav}(u,k) = \frac{\left[\sum_{v \in Co^{Fav}(u)} \frac{|\{q \in v.QF | I(q) \in I(k)\}|}{|v.QF|}\right]}{|Co^{Fav}(u)|} \quad (8)$$

We compute the similarities between k and activity interests of other users v who have co-watched, co-answered and co-favorited with a target user u in similar ways as shown in Equation 6, 7 and 8 respectively.

3 DATA EXTRACTION

There are two main datasets used in our study. For the GitHub dataset, we use the MongoDB database dump released on March 2015 [3]. The dataset contains GitHub activities from October 2013 to March 2015 of about 2.5 million users. Specifically, we are interested in the *fork* and *watch* repositories activities of the GitHub users. For Stack Overflow, we use the XML dataset released on March 2015[1]. This dataset contains information of estimated 1 million Stack Overflow users and their activities from October 2013 to March 2015. We are particularly interested in the *answer* and *favorite* activities of the Stack Overflow users.

User Account Linkage. As this study intends to predict user activities in GitHub and Stack Overflow, we need to identify users who were using both platforms. For this work, we used the dataset provided by Badashian et al. [1], where they utilized GitHub users' email addresses and Stack Overflow users' email MD5 hashes to find the intersection between the two datasets. We also filter out users who do not have at least 1 activity in both platforms between October 2013 and March 2015. In total, we identify 92,427 users, which forms our *base users* set. After the base users have been identified, we extract their GitHub and Stack Overflow activities from the datasets. In total, we have extracted 416,171 *fork*, 2,168,871 *watch*, 766,315 *answer* and 427,093 *favorite* activities.

Inferring User Interests To infer user topical interests in GitHub and Stack Overflow, we adopted the same heuristics proposed in [6], i.e, we infer the user topical interests in Stack Overflow and GitHub using the descriptive tags of the questions and repositories that they have participated.

4 EXPERIMENTS

In this section, we describe the supervised prediction experiments conducted to evaluate our proposed method. Specifically, we perform the following activity prediction tasks:

- *Answer Prediction.* Given a Stack Overflow *user-question* pair, predict if the user will answer the question
- *Favorite Prediction.* Given a Stack Overflow *user-question* pair, predict if the user will favorite the question
- *Fork Prediction.* Given a GitHub *user-repository* pair, predict if the user will fork the repository
- *Watch Prediction.* Given a GitHub *user-repository* pair, predict if the user will watch the repository

4.1 Experiment Setup

Data Selection. For *answer prediction* task, we retrieve all the Stack Overflow questions that the base users have answered and define a positive instance as a *user-question* pair where a base user had answered the particular question in Stack Overflow. For negative instances, we randomly assign a Stack Overflow question to the base users and check that the randomly assigned pair does not exist in the positive instance set. For the training datasets used in *answer prediction task*, we randomly generated 5,000 negative instances and randomly selected 5,000 positive instances from the questions answered by users between October 2013 and June 2014 (9 months). The same approach was used to generate the positive and negative instances for test sets using the questions answered by the users between July 2014 and March 2015 (9 months). Similar approach was used to generate the *user-question* and *user-repository* pairs for positive and negative instances used in *favorite, fork* and *watch prediction* tasks.

Feature Configuration. To compare the performance of *direct* and *cross* platform activity prediction approaches, we use Support Vector Machine (SVM) with linear kernel and apply the following feature sets on all prediction tasks:

- **SO_Act**: This set of features includes the *Answer* (Eqn. 3) and *Favorite* (Eqn. 4) *Interests Similarity* scores for a given user-question or user-repository pair.
- **SO_CoAct**: This set of features includes the *Co-Answer* (Eqn. 7) and *Co-Favorite* (Eqn. 8) *Interests Similarity* scores for a given user-question or user-repository pair.

[1] https://archive.org/details/stackexchange

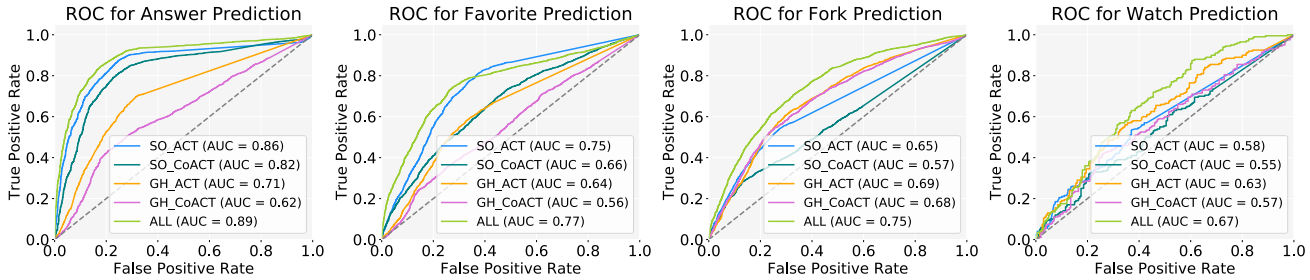

Figure 2: ROCs for Four Prediction Tasks

- **GH_Act**: This set of features includes the *Fork* (Eqn. 1) and *Watch* (Eqn. 2) *Interests Similarity* scores for a given user-question or user-repository pair.
- **GH_CoAct**: This set of features includes the *Co-Fork* (Eqn. 5) and *Co-Watch* (Eqn. 6) *Interests Similarity* scores for a given user-question or user-repository pair.
- **ALL**: This set of features is the union of all features.

4.2 Prediction Results

We measure the prediction accuracy for each feature configuration by computing the average area under the ROC curve (AUC) over a set of positive and negative examples drawn from the test set for each of the five runs. The results for the four prediction tasks are shown in Figure 2. We observe that feature configuration **ALL** performed the best in all prediction tasks, achieving an AUC of 0.89, 0.77, 0.75 and 0.67 for *answer*, *favorite*, *fork* and *watch* prediction tasks respectively.

Performance of cross platform prediction approach. Although the *cross platform prediction approach* did not outperform the *direct platform prediction approach* in user activity prediction, they still yield good accuracy. For example, when predicting user's *answer* and *favorite* activities in Stack Overflow, the GitHub *user activity interests similarity* features (i.e., **GH_Act**) has AUC of 0.71 and 0.64 respectively, and when predicting user's *fork* and *watch* activities in GitHub, the Stack Overflow *user activity interests similarity* features (i.e., **SO_Act**) has AUC of 0.65 and 0.58 respectively. The AUC for predicting user's answer activities in Stack Overflow using *user activity interests similarity* features (i.e., **GH_Act**) is observed to be slightly higher than the prediction for other activities. A possible explanation for this could be the difference between the nature of user activities; answering a question in Stack Overflow would require that a user possesses a particular domain expertise, whereas other activities such as watching a GitHub repository or favoriting a Stack Overflow question depend on the user's interests. As such, we observe higher AUC score for *predicting answer activity* task as the users' expertises are usually more specialized and less diverse than their interests.

More interestingly, using *cross platform prediction approach* with *user co-activity interests similarity* features (i.e., **GH_CoAct** and **SO_CoAct**), have also yielded reasonable prediction accuracies. For example, when predicting user's answer activities in Stack Overflow, **GH_CoAct** has yielded an AUC of 0.62. This suggests that even with no information about a user's past activities in the Stack Overflow and only minimal information such as the user's co-activities in GitHub, we are still able to reasonably predict user's activity in Stack Overflow. Similar observations are made when predicting user activities in GitHub using user's co-activities in Stack Overflow.

Solving cold-starts. The reasonably good accuracies of cross platform prediction approach also demonstrate its potential to solve cold-start problem; i.e., predicting and recommending a user's activities without knowing the users' past activity history in the platform. For example, when predicting user's answer activities in Stack Overflow, we are able to achieve AUC as high as 0.71 without using any Stack Overflow features (i.e, using GitHub features **GH_Act** only). Similar observations were made for fork, watch and favorite activities. We further conduct a small case study to retrieve and review fork predictions of users who did not have any past fork activities. For example, we successfully predicted that user *U420338* would forked repository *R12172473* in GitHub even when this was the first repository forked by the user (i.e., no past user fork activity). Examining into details, we found that *R12172473* has description tags ⟨*svg*, *javascript*⟩, and among the 95 questions *U420338* had answered in Stack Overflow, 83 contain the tags ⟨*javascript*⟩ or ⟨*svg*⟩ or both. By analyzing *U420338*'s Stack Overflow activities, our approach can identify his interests, which ultimately help in predicting the user's GitHub activities.

5 CONCLUSION AND FUTURE WORK

In this paper, we propose a novel framework which predicts users activities in multiple social collaborative platforms. We conducted experiments on large real-world datasets which contain activities of 92,427 users who are active in GitHub and Stack Overflow. Our proposed methods achieved good accuracy in predicting various user activities (up to an AUC score of 0.89). Our experiments shown that user activities in Stack Overflow can be predicted with reasonable accuracy using the same user's interests inferred from his or her activities in GitHub. The same observation was made when predicting a user's activities in GitHub using his or her interests inferred from his or her activities in Stack Overflow. The reasonable accuracies yield by cross platform prediction approach demonstrates its potential in solving cold-start problem in user activity prediction and recommendation in social collaborative platforms. For future work, we intend to consider more advance techniques (e.g., topic models or deep learning models) to derive and measure user interests similarity across multiple platforms. We will also consider platforms aside from Stack Overflow and GitHub (e.g. Quora).

REFERENCES

[1] Ali Sajedi Badashian, Afsaneh Esteki, Ameneh Gholipour, Abram Hindle, and Eleni Stroulia. 2014. Involvement, contribution and influence in GitHub and stack overflow. In *International Conference on Computer Science and Software Engineering (CCSE)*.

[2] Lucas B. L. de Souza, Eduardo C. Campos, and Marcelo de A. Maia. 2014. Ranking Crowd Knowledge to Assist Software Development. In *International Conference on Program Comprehension (ICPC)*.

[3] Georgios Gousios. 2013. The GHTorrent dataset and tool suite. In *International Conference on Mining Software Repositories (MSR)*.

[4] Mohamed Guendouz, Abdelmalek Amine, and Reda Mohamed Hamou. 2015. Recommending relevant GitHub repositories: a collaborative-filtering approach. *on Networking and Advanced Systems* (2015).

[5] Jyun-Yu Jiang, Pu-Jen Cheng, and Wei Wang. 2017. Open Source Repository Recommendation in Social Coding. In *SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

[6] Roy Ka-Wei Lee and David Lo. 2017. GitHub and Stack Overflow: Analyzing Developer Interests Across Multiple Social Collaborative Platforms. In *International Conference on Social Informatics (SocInfo)*.

[7] Fatemeh Riahi, Zainab Zolaktaf, Mahdi Shafiei, and Evangelos Milios. 2012. Finding Expert Users in Community Question Answering. In *International World Wide Web Conference (WWW)*.

[8] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. 2002. Methods and metrics for cold-start recommendations. In *SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

[9] Giuseppe Silvestri, Jie Yang, Alessandro Bozzon, and Andrea Tagarelli. 2015. Linking Accounts across Social Networks: the Case of StackOverflow, Github and Twitte.. In *International Workshop on Knowledge Discovery on the WEB*.

[10] Bogdan Vasilescu, Vladimir Filkov, and Alexander Serebrenik. 2013. StackOverflow and GitHub: associations between software development and crowdsourced knowledge. In *International Conference on Social Computing and Networking (SocialCom)*.

[11] Lin Wang, Bin Wu, Juan Yang, and Shuang Peng. 2016. Personalized recommendation for new questions in community question answering. In *International conference series on Advances in Social Network Analysis and Mining (ASONAM)*.

[12] Wei Wang, Haroon Malik, and Michael W. Godfrey. 2015. Recommending Posts Concerning API Issues in Developer Q&sA Sites. In *International Conference on Mining Software Repositories (MSR)*.

[13] Congfu Xu, Xin Wang, and Yunhui Guo. 2016. Collaborative Expert Recommendation for Community-Based Question Answering. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMP-PKDD)*.

[14] Liu Yang, Minghui Qiu, Swapna Gottipati, Feida Zhu, Jing Jiang, Huiping Sun, and Zhong Chen. 2013. CQArank: Jointly Model Topics and Expertise in Community Question Answering. In *International Conference on Information and Knowledge Management (CIKM)*.

[15] Yue Yu, Huaimin Wang, Gang Yin, and Charles X Ling. 2014. Reviewer recommender of pull-requests in GitHub. In *International Conference on Software Maintenance and Evolution (ICSME)*.

[16] Lingxiao Zhang, Yanzhen Zou, Bing Xie, and Zixiao Zhu. 2014. Recommending relevant projects via user behaviour: an exploratory study on github. In *Workshop on Crowd-based Software Development Methods and Technologies (CrowdSoft)*.