基于 GitHub 用户行为的开发者社团发现

开题汇报

孙秋实 宋晏如

数据科学与工程学院 2018 级

March 25, 2022

Outline

项目背景

挖掘 GitHub 潜在的社交属性，通过社团发现来辅助挖掘其社交潜力

“物以类聚，人以群分”

1. “物” → Repository
2. “人” → Contributor

在社团中，进一步挖掘贡献者/仓库的合作关系

社团内部

1. Contributors: Follow, Star...
2. Repository: Fork, PR...

- 找出互相刷"star" 的水军（反作弊）
- 找出协作紧密且适合社交的组织（挖掘社交）

研究方法：图算法

暂定使用以下两种算法对 GitHub 用户行为和仓库进行社团发现

谱方法

- 构建模块度矩阵
- 迭代式社区分裂

Louvain 算法

- 自底向上合并小社区

研究特点

对社交潜力的挖掘还需要了解使用者的行为习惯，除了对稀疏图本身性质的研究与应用之外，我们还需要将以下因素纳入考虑。

纳入分析的因素

- 时空数据
  - Contributor 的 GitHub 活跃时间段与活跃度
  - Contributor 所在的地理位置（语言）
- 使用者的身份（如员工身份）
- ...

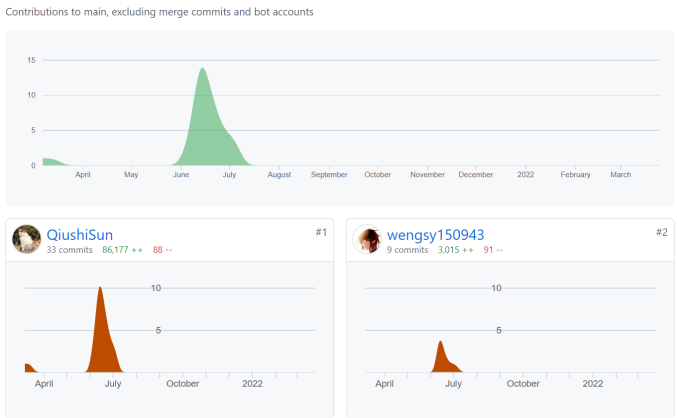

Figure: 仓库的活跃度例子

项目计划

项目计划

1. GitHub 数据集的预处理
2. 图构建（利用 Neo4j 等图数据库）
3. 基于用户行为与用户特征的社团发现[a]
4. 对社团发现效果进行评估，比较上述算法在该问题的性能差异
5. 社团关系的可视化

[a]若有余力，对重叠社团进行进一步分析

Thank You!