基于 GitHub 用户行为的开发者社团发现

孙秋实 宋晏如

数据科学与工程学院 2018 级

2022. 4. 15

目录

- 数据预处理
- 图的构建
- 近期计划

数据读取

- 从 log 中读取;
- 读取属性：{id, type, actor_id, actor_login, repo_id, repo_name, org_id, org_login, pt};
- 筛选了 2020 年 1 月 1 号及之后的操作。

数据整理

- actor_info: {actor_id, actor_login},
- repo_info: {repo_id, repo_name},
- org_info: {org_id, org_login},
- actor_org: {actor_id, org_id},
- action_info: {id, type, actor_id, repo_id, pt }.

数据清洗

- 只保留对活跃 repo 的操作;
- 只保留非 bot 的 actor 的操作。

图的构建

- 以 actor 为节点;
- 将每个用户对所有 repo 分别的操作数构造为向量;
- 对于有共同操作 repo 的用户，计算向量相似度取倒数作为距离。

近期计划

- 读取数据、获取不同操作的详细信息，并进行处理与分析;
- 完善 repo 与 actor 筛选算法;
- 修改图的节点之间距离的算法;
- 对 repo 之间的协作关系进行可视化。

谢谢！