# 基于加权PageRank的Github项目活跃度分析

## 1.数据预处理

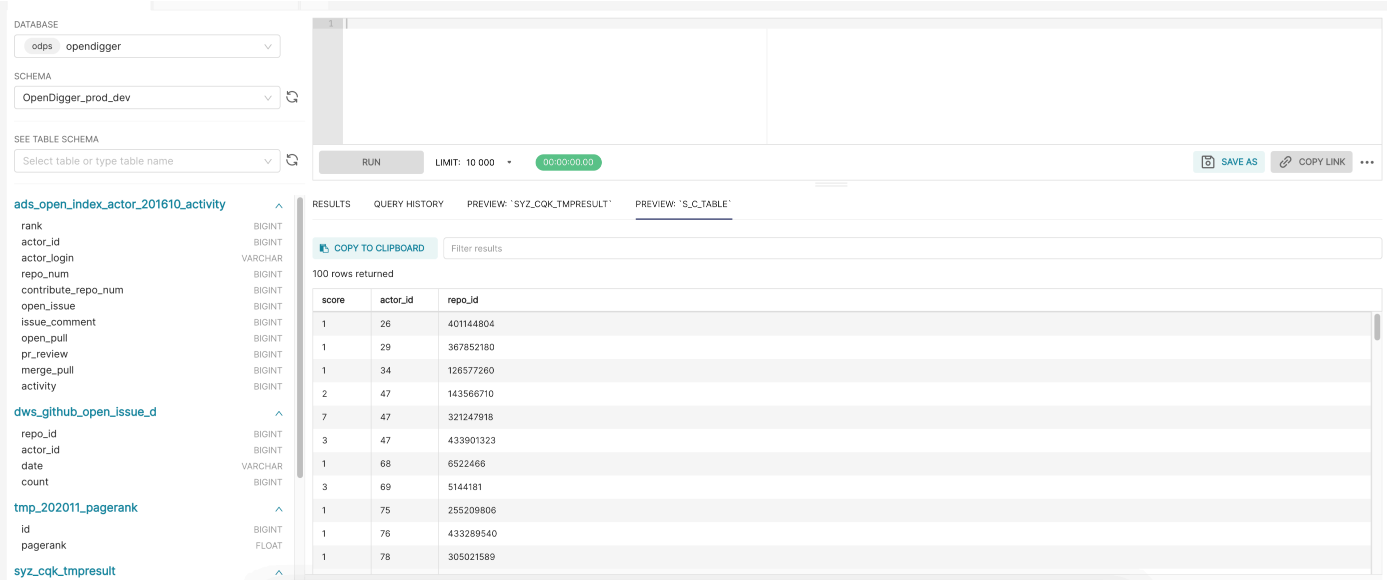
### 1.1基于开发者的 GitHub 行为数据计算其对项目的贡献度

$A_d = \sum w_i c_i$ ( 其中的 $A_d$ 为开发者对项目的贡献度，而 $c_i$为行为事件由该开发者触发的发生次数，$w_i$ 为该行为事件的加权比例)

| 行为事件 | 赋分 |
|---|---|
| Issue 评论 | 1 |
| Issue | 2 |
| PR | 3 |
| PR Review | 4 |

```sql
SELECT  SUM(t.score) AS score, actor_id, repo_id
FROM
  (SELECT CASE WHEN type = 'IssueCommentEvent' THEN 1
   WHEN type = 'IssuesEvent' THEN 2 when  type = 'PullRequestEvent' THEN 3
   WHEN type = 'PullRequestReviewCommentEvent' THEN 4
   ELSE 0 END AS score ,actor_id,repo_id
   FROM ods_github_log
   WHERE pt="20211201" ) t
GROUP BY actor_id,repo_id
```

将统计好的数据存入中间表s_c_table中：



DATABASE: odps opendigger

SCHEMA: OpenDigger_prod_dev

SEE TABLE SCHEMA

LIMIT: 10 000

00:00:00.00

ads_open_index_actor_201610_activity

| | |
|---|---|
| rank | BIGINT |
| actor_id | BIGINT |
| actor_login | VARCHAR |
| repo_num | BIGINT |
| contribute_repo_num | BIGINT |
| open_issue | BIGINT |
| issue_comment | BIGINT |
| open_pull | BIGINT |
| pr_review | BIGINT |
| merge_pull | BIGINT |
| activity | BIGINT |

dws_github_open_issue_d

| | |
|---|---|
| repo_id | BIGINT |
| actor_id | BIGINT |
| date | VARCHAR |
| count | BIGINT |

tmp_202011_pagerank

| | |
|---|---|
| id | BIGINT |
| pagerank | FLOAT |

syz_cqk_tmpresult

100 rows returned

| score | actor_id | repo_id |
|---|---|---|
| 1 | 26 | 401144804 |
| 1 | 29 | 367852180 |
| 1 | 34 | 126577260 |
| 2 | 47 | 143566710 |
| 7 | 47 | 321247918 |
| 3 | 47 | 433901323 |
| 1 | 68 | 6522466 |
| 3 | 69 | 5144181 |
| 1 | 75 | 255209806 |
| 1 | 76 | 433289540 |
| 1 | 78 | 305021589 |

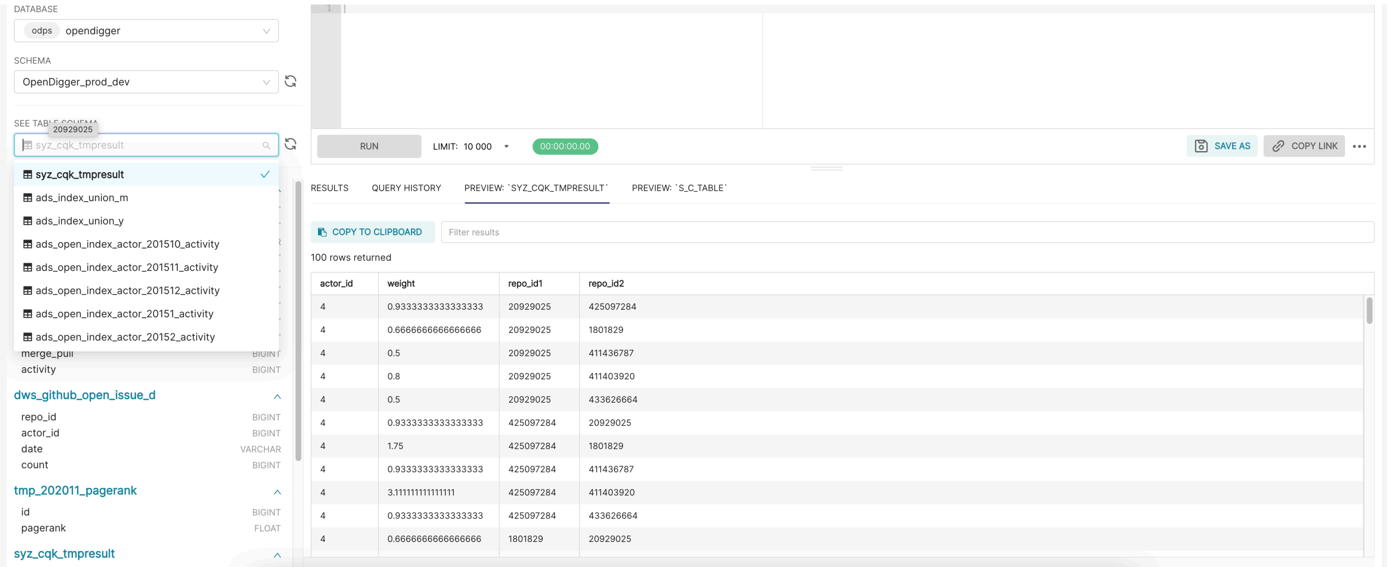
## 2.构建开源协作网络

### 筛选数据：

```
SELECT *
FROM s_c_table
WHERE actor_id not in
(SELECT actor_id FROM s_c_table GROUP BY actor_id HAVING COUNT(*) = 1)
```

### 计算项目与项目之间的协作关联度：

```
select s1.actor_id, s1.score*s2.score/(s1.score+s2.score) as weight, s1.repo_id  AS
repo_id1,s2.repo_id AS repo_id2
FROM syz_cqk_tmptable s1 JOIN syz_cqk_tmptable s2
ON s1.actor_id = s2.actor_id AND s1.repo_id != s2.repo_id
```



DATABASE: odps opendigger

SCHEMA: OpenDigger_prod_dev

SEE TABLE SCHEMA: syz_cqk_tmpresult

- syz_cqk_tmpresult ✓
- ads_index_union_m
- ads_index_union_y
- ads_open_index_actor_201510_activity
- ads_open_index_actor_201511_activity
- ads_open_index_actor_201512_activity
- ads_open_index_actor_20151_activity
- ads_open_index_actor_20152_activity

| | |
|---|---|
| merge_pull | BIGINT |
| activity | BIGINT |

dws_github_open_issue_d

| | |
|---|---|
| repo_id | BIGINT |
| actor_id | BIGINT |
| date | VARCHAR |
| count | BIGINT |

tmp_202011_pagerank

| | |
|---|---|
| id | BIGINT |
| pagerank | FLOAT |

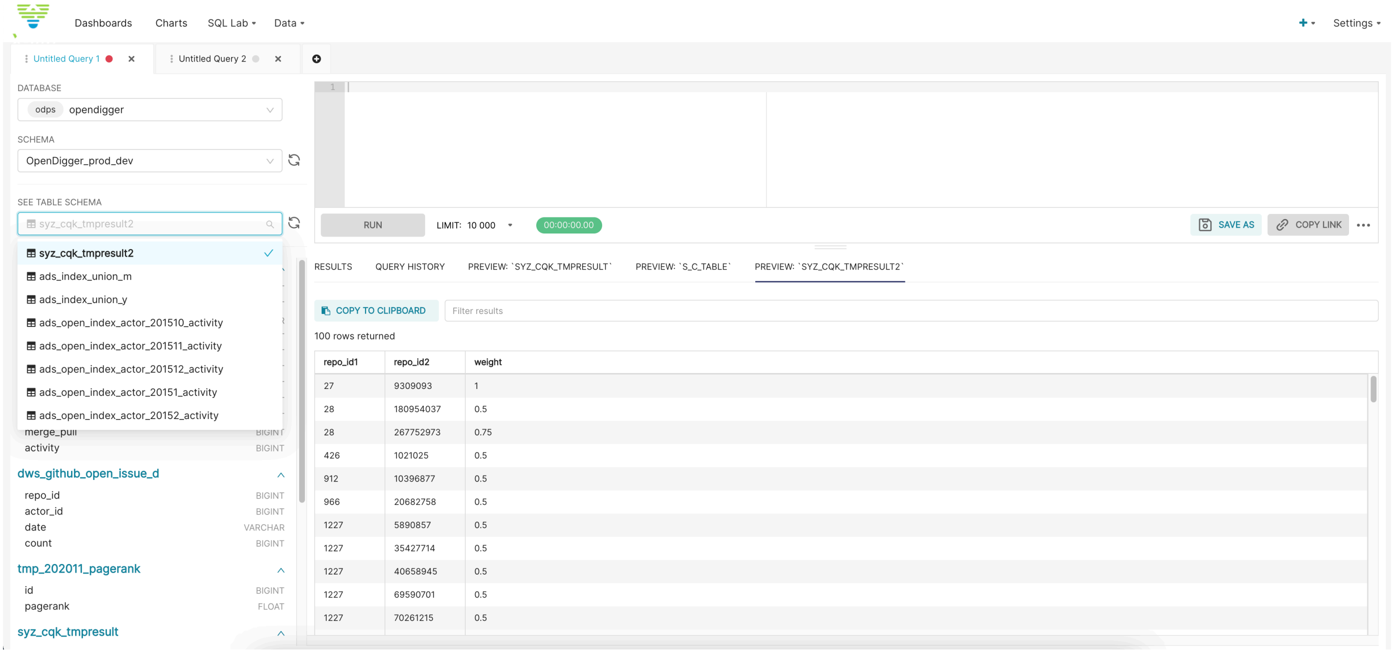
syz_cqk_tmpresult

100 rows returned

| actor_id | weight | repo_id1 | repo_id2 |
|---|---|---|---|
| 4 | 0.9333333333333333 | 20929025 | 425097284 |
| 4 | 0.6666666666666666 | 20929025 | 1801829 |
| 4 | 0.5 | 20929025 | 411436787 |
| 4 | 0.8 | 20929025 | 411403920 |
| 4 | 0.5 | 20929025 | 433626664 |
| 4 | 0.9333333333333333 | 425097284 | 20929025 |
| 4 | 1.75 | 425097284 | 1801829 |
| 4 | 0.9333333333333333 | 425097284 | 411436787 |
| 4 | 3.111111111111111 | 425097284 | 411403920 |
| 4 | 0.9333333333333333 | 425097284 | 433626664 |
| 4 | 0.6666666666666666 | 1801829 | 20929025 |

汇总：

```
select repo_id1, repo_id2, SUM(weight) as weight
from SYZ_CQK_TmpResult
Group by repo_id1,repo_id2
```



DATABASE: odps opendigger

SCHEMA: OpenDigger_prod_dev

SEE TABLE SCHEMA: syz_cqk_tmpresult2

- syz_cqk_tmpresult2 ✓
- ads_index_union_m
- ads_index_union_y
- ads_open_index_actor_201510_activity
- ads_open_index_actor_201511_activity
- ads_open_index_actor_201512_activity
- ads_open_index_actor_20151_activity
- ads_open_index_actor_20152_activity

| | |
|---|---|
| merge_pull | BIGINT |
| activity | BIGINT |

dws_github_open_issue_d

| | |
|---|---|
| repo_id | BIGINT |
| actor_id | BIGINT |
| date | VARCHAR |
| count | BIGINT |

tmp_202011_pagerank

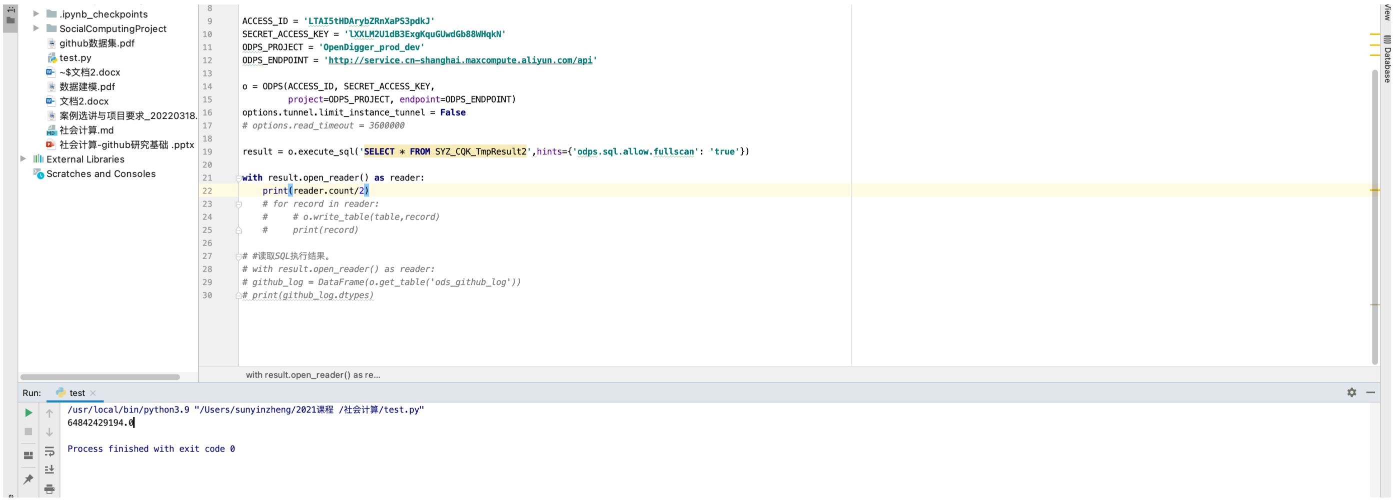
| | |
|---|---|
| id | BIGINT |
| pagerank | FLOAT |

syz_cqk_tmpresult

100 rows returned

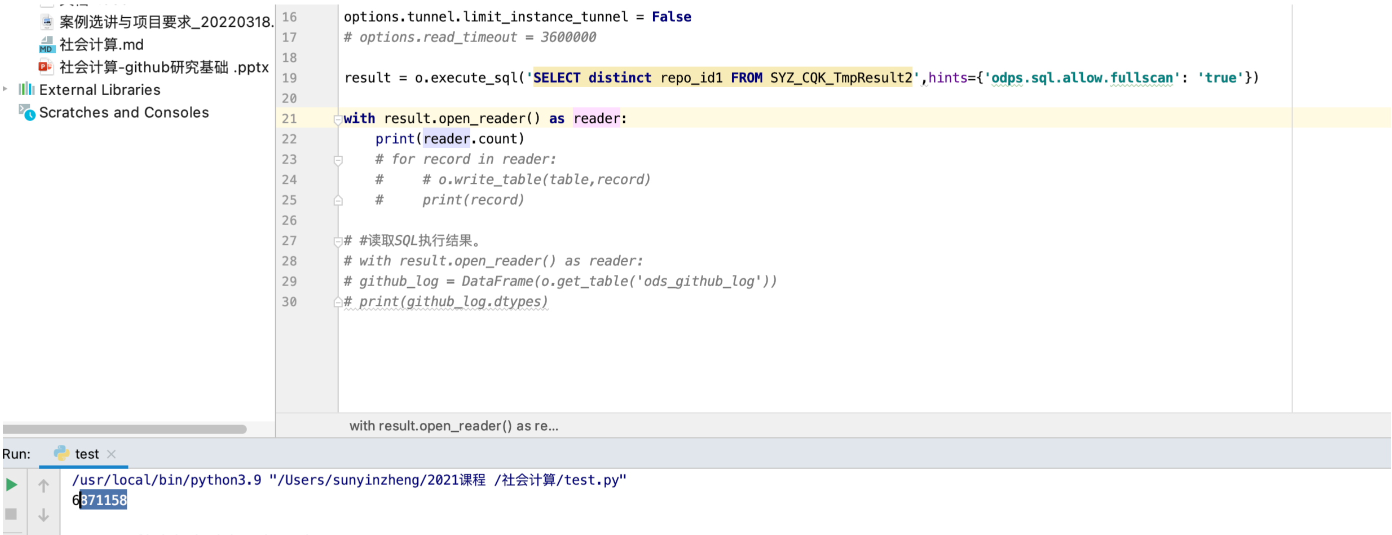
| repo_id1 | repo_id2 | weight |
|---|---|---|
| 27 | 9309093 | 1 |
| 28 | 180954037 | 0.5 |
| 28 | 267752973 | 0.75 |
| 426 | 1021025 | 0.5 |
| 912 | 10396877 | 0.5 |
| 966 | 20682758 | 0.5 |
| 1227 | 5890857 | 0.5 |
| 1227 | 35427714 | 0.5 |
| 1227 | 40658945 | 0.5 |
| 1227 | 69590701 | 0.5 |
| 1227 | 70261215 | 0.5 |

边的条数：648亿 顶点个数: 687万



```python

ACCESS_ID = 'LTAI5tHDArybZRnXaPS3pdkJ'
SECRET_ACCESS_KEY = 'lXXLM2U1dB3ExgKquGUwdGb88WHqkN'
ODPS_PROJECT = 'OpenDigger_prod_dev'
ODPS_ENDPOINT = 'http://service.cn-shanghai.maxcompute.aliyun.com/api'

o = ODPS(ACCESS_ID, SECRET_ACCESS_KEY,
         project=ODPS_PROJECT, endpoint=ODPS_ENDPOINT)
options.tunnel.limit_instance_tunnel = False
# options.read_timeout = 3600000

result = o.execute_sql('SELECT * FROM SYZ_CQK_TmpResult2',hints={'odps.sql.allow.fullscan': 'true'})

with result.open_reader() as reader:
    print(reader.count/2)
    # for record in reader:
    #     # o.write_table(table,record)
    #     print(record)

# #读取SQL执行结果。
# with result.open_reader() as reader:
# github_log = DataFrame(o.get_table('ods_github_log'))
# print(github_log.dtypes)
```

```
/usr/local/bin/python3.9 "/Users/sunyinzheng/2021课程 /社会计算/test.py"
64842429194.0

Process finished with exit code 0
```


案例选讲与项目要求_20220318.
社会计算.md
社会计算-github研究基础 .pptx
External Libraries
Scratches and Consoles
with result.open_reader() as re...
Run:
test

```python
options.tunnel.limit_instance_tunnel = False
# options.read_timeout = 3600000

result = o.execute_sql('SELECT distinct repo_id1 FROM SYZ_CQK_TmpResult2',hints={'odps.sql.allow.fullscan': 'true'})

with result.open_reader() as reader:
    print(reader.count)
    # for record in reader:
    #     # o.write_table(table,record)
    #     print(record)

# #读取SQL执行结果。
# with result.open_reader() as reader:
# github_log = DataFrame(o.get_table('ods_github_log'))
# print(github_log.dtypes)
```

```
/usr/local/bin/python3.9 "/Users/sunyinzheng/2021课程 /社会计算/test.py"
6871158
```

## 3.后续计划

通过pagerank算法来计算项目在开源协作网络中的影响力。