# Python图算法江西银行现场验证

目录

[Python图算法江西银行现场验证 1](#_Toc8599)

[一、背景 2](#_Toc11751)

[二、验证流程 2](#_Toc26706)

[三、数据准备 2](#_Toc22835)

[四、权重计算 2](#_Toc11159)

[五、数据入图 3](#_Toc29955)

[六、算法选择 3](#_Toc2025)

[七、结果保存 3](#_Toc12010)

[八、结果分析 3](#_Toc21310)

[九、其他 4](#_Toc12528)

[十、应用与展望 4](#_Toc13246)

## 一、背景

基于前半年调研并开发了基于python实现的图算法，正赶上有本次江西银行的poc测试项目，所以考虑利用poc测试的闲暇时间，验证一下图算法能力。

## 二、验证流程

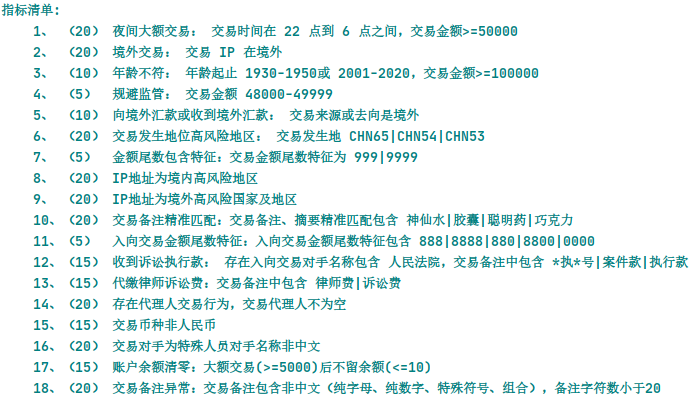
数据准备-->权重计算-->数据入图-->调用算法-->保存结果-->分析结果

## 三、数据准备

基于poc行方给的所有交易流水数据，原始流水量为1.23亿。经过滤第三方账户及一人多卡互转等无效交易流水数据后，流水量还有大概2000w左右。最后落成表，做为图算法的数据输入。

## 四、权重计算

大体策略是利用专家规则剥离成针对单条交易流水的特征指标来计算权重，命中则给予对应的规则分数。最后先求出单条交易流水命中规则的总分数，再对总分数做归一化缩放。做为图上边的权重。详情单流水交易特征指标参考图4.1所示。



**图 4.1**

## 五、数据入图

初始化无向多边图

## 六、算法选择

因时间、环境等问题。故在现场只调通了LOUVAIN算法，并通过调整算法超参数，最终跑了两轮的结果输出。

## 七、结果保存

## 八、结果分析

综合两次算法输出结果，通过随机抽查团伙复核后，总结得出如下结论：

1. 算法筛选出的团伙中有一个或两个账户为政府、医院、超市等账户。此类账户虽然在交易行为上仅可以认为是一个社区，但不能说是一个团伙。后续需要考虑通过关键字或白名单等方式来过滤掉此类账户。
2. 可能是因为图是无向图，权重没有更好的发挥作用。
3. 数据入图前进行过滤，比如设定一个阈值，将计算后的权重值小于此阈值的流水再次进行过滤，以大大减少上图的数据量。
4. 输出的团伙也可以进行优化，因为发现算法输出结果中包含一些2-3个账户的团伙。所以可以将团伙中账户个数少于2-3个的团伙结果过滤掉。
5. ……

## 九、其他

关于资源使用情况，流水2000w数据跑一整个图算法，内存占用最高时达到40G（初步判断大概率是在计算权重部分占用）。

资源优化想法：

1. 改权重计算为sql计算并落表到数据库。
2. ……

## 十、应用与展望

目前来看，图算法项目还有很大的优化空间，不论是从数据层面还是算法层面来说还需要进一步完善运行机制。例如前期数据建设方面，白名单、三方关键字的收集与过滤。而在算法层面则需要完善的是如何在计算权重能减少内存使用、又如何减少上图数据以更好的辅助以减少权重计算所使用的内存。因此次poc时间有限，像团伙评估指标，紧密度、中心度还未来得及验证。

关于后续如何应用，图算法可以提供api接口，由甄别平台调用，得到结果后在平台展示。只要保证后续项目数据库表版本一致，可无修改应用于项目中。