2022 年中述职

反爬业务 BP 组 王童亮 wangtongliang 2022 年 7 月

一、总结过去

1. 目标与结果

业务目标：22 年整体，优选反爬目标以 case 个数衡量：S1≤1、S2≤1、S3≤5、S4≤5，截至目前，有 1 个 S1 和 5 个 S9。

上半年，通过三件套确保视野齐全、通过运营 SOP 一定程度上重新定义“具备能力”，case 个数和规模显著下降，但还无法承诺稳定收敛的趋势，主要原因是对反爬能力的缺乏客观、量化的描述，不能保障：某规模、某类型的爬虫一定防得住。所以：标准化的反爬能力、标准的运营流程是下个阶段的重点工作，讲白话就是希望能把反爬的能力水位客观地定义清楚，即应对特定类型的爬虫手法，在各个环节分别具备&欠缺哪些能力，when&how 补齐，能拦截什么水平、什么规模的爬虫。

项目目标：

1）三件套：找全接口、都登录、都 log 的继定目标基本达成，累计治理未登录接口 20+（登录覆盖率 100%），未 log 接口 300+（log 覆盖率 96%）。从结果看，今年以来，因视野问题导致的爬虫漏过仅有 1 例，且是通过三件套发现。虽然项目目标顺利达成，但是由于缺乏项目管理经验，导致实现路径不清晰，达成过程不顺利。

2）运营 SOP：组织搭建了优选 mvp 版本的运营方案，对反爬全景的全部能力的输出提供了一组观测指标，把在已有能力下发现大规模爬虫从偶然变成必然。自 5 月份建成以来，累计发现 2 例 case（召回 33%）和 3 例爬虫风险，但目前能力很不成熟，一是只能发现能力范围内的大规模的爬虫，二是对只能能力的输出指标，对能力本身质量和覆盖情况没有保障，标准化需要提上日程。

3）反跟价：确定了竞对真人现看现跟一轮的跟价模式，已经封号实验基本论证了反爬可以影响 lose 率的事实，下一步的计划是通过自动地持续封禁，具备压制 lose 率的能力。并且通过建配套指标、规范分析思路，保证对竞对跟价模式的持续监控和对短期上涨的快速响应。

2. 分析与总结

1）**对优选反爬整体复盘**：结果向好，case 个数和规模都在收敛，低级错误导致的大规模 case 基本消失。

做得好的地方：

关键内因是正确的指导思想：**保全局视野、抓主要矛盾**，主要对应下方两个具体的项目。

A）定义反爬全景，确定风险预防的主要矛盾，大搞三件套，重点解决；

B）建运营 SOP，做好风险兜底，先保证不犯低级错误，不漏大规模爬虫；

客观地讲，外因可能是对手在进攻方向的目标放低，投入减少，没有那么多爬虫给我们防，case 数自然减少。

附：反爬全景&最近半年的提升项

方向	能力	半年前	现在	做的事情
风险预防	找全接口	★★☆	★★★★☆	完善了扫描方案
	都登录	★★☆	★★★★☆	建成扫描能力，完成存量治理
	都 log	★★	★★★★	业务上报+反爬解析
	能力覆盖率	★	★★	人工梳理覆盖策略
风险感知	业务指标	★	★★★	引入 lose 率、情报等结果指标
	异常报出	★	★★	输出结果每周运营
风险识别	行为	★	★★☆	累计维度、周期 MECE 分类
	端	★★	★★	/
	账号	★★★	★★★	/

做得不好的地方：

A) 没有项目管理意识和能力，做事情没有章法。主要还是靠项目成员的竭力支持在保证产出，属于“卷”出来的成果，效率和质量都没有保障，虽然结果向好，但是过程都比较坎坷。

B) 反爬能力的判断标准太随意。对具备能力的定义过于单薄，甚至排查一次存量发现没有爬虫就判断具备能力，说好听点叫乐观，说难听点就是不负责，接受了太多模棱两可的“DONE”，导致不犯重复错误这个基本要求也难以达成。

二、展望未来

***1.* 优选反爬的整体规划：**目标是稳定可靠，手段是反爬能力标准化，方案是给全部能力建一套客观的评价指标。在没 *case* 是常态的情况下，具备说清：当前能力水位可以保障不出 *xx* 规模 *case* 的能力，并且可以靠蓝军测试对结论进行佐证。

能力按照爬虫手法拆解，评价指标包括三个部分：能力质量、能力覆盖程度、能力输出结果的运营情况。

下图是从 *BP* 视角出发，对能力现状做了感性评估，确定了部分方向的指标思路以及下个阶段的理想目标。

方向	任务	现状	指标	目标
覆盖	找全接口	★★★★☆	*1*、准确率 *2*、召回率	稳定保障全部关键数据接口都登录、都 *log*
	都登录	★★★★☆	*1*、登录覆盖率 *2*、不登录流出的数据量级	
	都 *log*	★★★★	*1*、*log* 覆盖率 *2*、*log* 完备率 *3*、分类的触底召回率	
	验签	★★★☆	*1*、*Now-1* 版本的验签覆盖率 *2*、攻破验签的对手个数	高强度的新版本验签 *100%*覆盖
端防	冗余上报	☆	*1*、多源校验参数覆盖率	提高竞对的改机成本
	异常环境	★★	*1*、高风险标签的拦截情况 *2*、中低风险标签的运营情况	
	生物探针	☆	*1*、数据覆盖率 *2*、输出结果的运营情况 *3*、能力本身的准召	
账号	黑灰号	★★	*1*、标签本身的准召 *2*、黑号的拦截情况 *3*、灰号的运营情况	账号成本显著 *beat* 竞对
	token 安全性	★★★☆	*1*、校验策略的完备性	合法美团账号才能访问
	新即原罪	★	*1*、纯新号的爬虫规模 *2*、新号输出结果的运营情况	攻击者搞不到大规模的新号
	微信账号校验	★★★★☆	*1*、校验策略的完备性	合法微信账号才能访问
行为	各维度访问限制	★★☆	*1*、极限情况单账号可获取的数据量级（*1d/7d/15d/30d*） *2*、策略的 *MECE* 程度描述（累计维度、覆盖场景等） *3*、卡阈值账号的运营情况	单账号获取的数据量级有限
	爬虫 *SOP* 打击	★★	*1*、满足 *SOP* 定义的账号量级 *2*、*SOP* 对跟价/*case* 的召回	爬虫需要伪造自己的行为才能正常获数
	低信誉账号降级	★	*1*、低信誉账号定义的准召 *2*、低信誉账号的量级 *3*、单位爬虫账号的养号成本	爬虫需要养号来获取高质量号源
IP	黑灰 *IP* 打击	★	*1*、异常 *IP* 输出结果的运营情况	大规模的异常 *IP* 爬虫能被及时感知
运营	指标运营	★★	*1*、能力*指标的组合个数 *2*、健康的指标个数 *3*、有运营 *SOP* 保障的指标个数	反爬能力都能通过指标体现，爬虫 *case* 可以通过分析指标波动发现
	策略覆盖	★★☆	*1*、各个方向涉及的策略个数 *2*、策略*场景*接口的覆盖率	既有能力做到 *100%*覆盖

***2.* 对反跟价的规划：**

务实地看：尽快上线对竞对身份*行为 *SOP* 的自动封号，目的是封全封准的同时不让竞对认为物理身份是封禁根因，以保持对跟价账号的持续追踪和限制，以此压制 *lose* 率。

务虚地看：希望跟竞对达成一种制衡的状态，尝试找到一种达成我方目标但又不“惹恼”竞对的方案，前提条件是双方的目标不绝对冲突，有空可钻（比如关注的时间等），但具体怎么做还没想法。