基于图算法的跨设备受众识别

刘喆 liuzhe@admaster.com.cn

目录

- 问题是什么
- AdMaster 的方案
- 工程实现和权衡
- 有趣的副产品

问题是什么

- 一人多机
- 有固定账号体系是一种幸福
  - QQ
  - taboo
  - JD
  - 天涯
- 没有账号体系怎么办?
  - cookie
  - deviceID

问题是什么

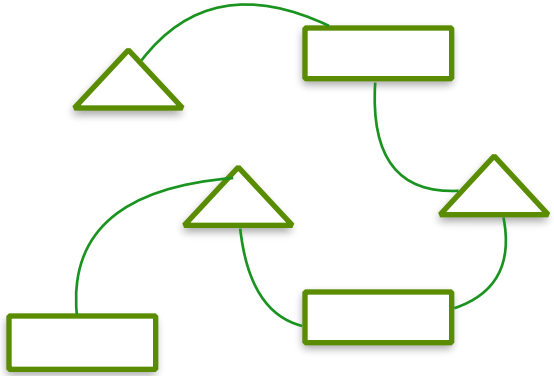
- cookie 的攻防战
  - 隐身模式
  - “安全”浏览器
  - 定期清 cookie
- DeviceID 的尴尬
  - idea (md5)
  - mac (aa:bb:cc:dd:ee aabbccddee md5 AABBCCDDEE MD5)
  - IMEI (15 位, 14 位, md5, MD5)
  - openudid
- 统一第三方 id
  - googleID
  - qq

AdMaster 的方案

- 假设: 人, 一定会在同一个地方用不同的设备上网.
- 奇葩: 只要在不同的地方, 就用不同的设备上网
  - 我只在网吧上网
  - 我在家用家里的电脑, 在公司用公司的电脑, 不用手机
  - 我很长一段时间内(比如 4 周), 只用电脑, 另一段时间只用手机
- 不同的人, 用同一个设备
  - 我和老婆共用一个电脑
  - 一家人共用一台电视 (tv 到人识别, 另一个话题)

AdMaster 的方案

- 我们认为:
  - 有多个设备的人, 一定会用电脑
  - 所以一定可以用 cookie - device 这样的关联来联系起来



AdMaster 的方案

- 每次取 15 天的访问数据 (?)
- 同一天, 同一 ip, 找出 cookie 和 device, 做笛卡尔积, 形成 pair (?)
- 每个 pair 对生成特征向量
- 根据已知数据, 对特征向量训练, 生成模型
- 把模型应用到新的 pair, 根据可信度, 取可信的 pair
- 用这些 pair 生成森林, 每棵树就是同一个人

工程实现和权衡

- 特征取哪些
  - cookie - ip - count
  - device - ip - count
  - cookie-ips　device-ips　common count
  - ip-public-weight
    - sum(pair-count / ip-public-weight)　　越大越好
  - same web page count
  - 同一电视剧

工程实现和权衡

- 清洗
  - blueAir 10%
  - 15 天内, 同一个 ip 出现的次数应该小于 4000 次 40%
- 训练
  - gbdt
  - xgboost with spark

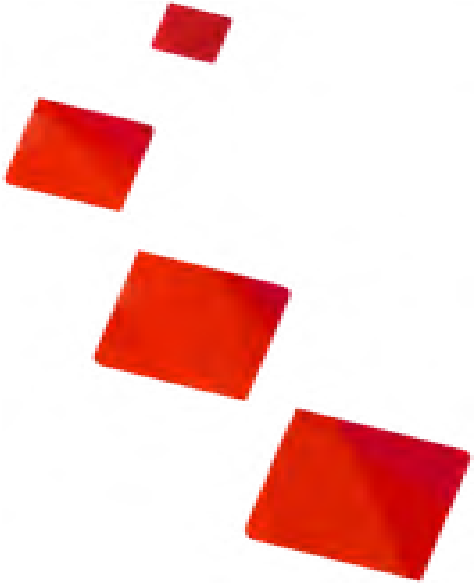
工程实现和权衡

- 森林生成算法
  - GraphX
- 基于 aerospike 自己实现
  - kv 数据库
  - 4 台 机器 160w/s
  - 把 pair (C-D)当成流, 逐个加入, 用两张表, 一个是 key-superID, 另一个是, superID-keys

两个 *key* 都不在库里	两个都加入库里, 共用同一个 *superID*
只有一个 *key* 在库里	把不在库里的, 加入库里, 用另一个的 *superID*
两个 *key* 都在库里	*superID* 合并, 把另一个*superID* 下所有的 *keys* 的*id* 修改

有趣的副产品

- 群组大小最多的, 3-7, 似乎不合逻辑(人手两个手机, 一台电脑, 基本上标准配置了, 3 个也足够了), 原因在于, 不同数据源得到的数据格式是不相同的, 以 android 为例
  - 32 位 imei md5 大写是 MMA 的标准做法
  - 15 位 imei 原值
  - 14 位 imei, 没有校验吗
  - 32 位 imei md5 后再 md5
- 可以发现很多作弊的 id
  - 最大的群组大小为 267, 同一个设备id, 不同的 cookie
  - 是他只是简单的清 cookie 了吗? 不是的, 它的访问事件, 在每天的每个小时都有, 总不可能这人不睡觉一直在上网吧?



Q & A

QQ 30592378
liuzhe@admaster.com.cn