

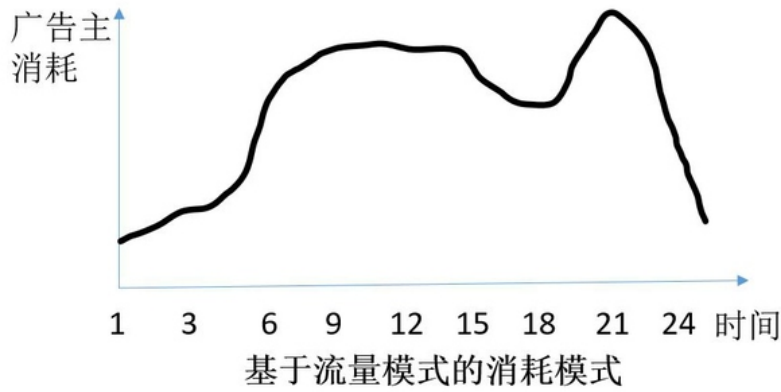
广告预算平滑消耗调研

张航

May 2017

1 方法简介

奇效广告平滑消耗是指将广告系列单日的预算按一定时间粒度进行平滑，避免广告集中曝光。平滑消耗可以帮助广告主触及更多不同用户，进而提高转化率。常用的平滑方式基于流量在一天中的分布，如下图所示，每个时段广告消耗应该正比于流量大小，我们也将采用这种方式。



由于预算设置在广告系列 (order group), 我们的平滑消耗也在广告系列粒度。如果已知系列 i 的总预算 B_i , 截至 t 时刻已消耗预算 $B_i(t)$, 消耗比例 $b_i(t) = B_i(t)/B_i$, 则平滑消耗需要给出 $p_i(t)$, 即 t 时刻系列 i 下的订单通过率 (pass-through rate, PTR), 满足

$$p_i(t) = \begin{cases} \min((1 + r_t) \times p_i(t-1), 1) & \text{if } b_i(t) \leq (1 + \alpha)c_i(t) \\ \max((1 - r_t) \times p_i(t-1), 0) & \text{if } b_i(t) > (1 + \alpha)c_i(t) \end{cases} \quad (1)$$

其中 $c_i(t)$ 为流量截至 t 时刻的累积分布。为了加速消耗, 尽可能避免平滑影响收入, 我们对目标比例添加了加速因子 α , 设 $\alpha = 0.1$, 若 t 时刻, 正常消耗比例是 50%, 加速消耗比例为 55%。结合奇效广告系统的特点, 要给出 $p_i(t)$ 需要解决两个问题:

1. 订单项 (order plan) 有定向条件, 不同定向条件对应的流量分布可能差异较大, 并且定向条件可能会被频繁修改;

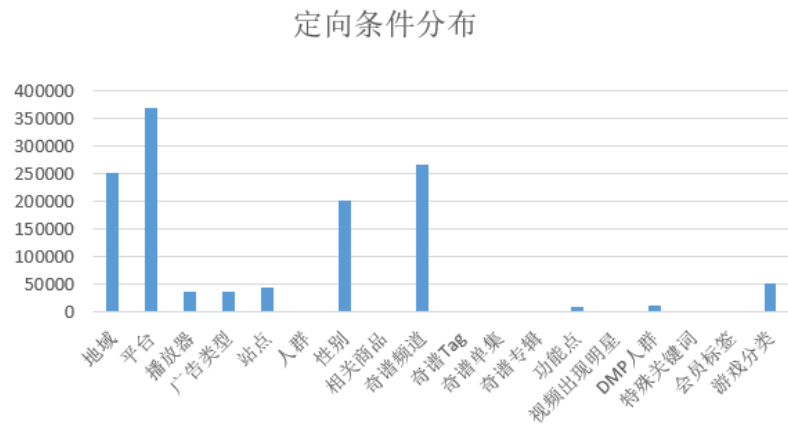


图 1: 有某种定向条件的订单项个数，如有 25 万订单包含地域定向

2. 预算设置在广告系列 (order group)，而一个广告系列可能包含多个订单项，分别定向到不同广告位，它们的曝光次数，eCPM 等差异可能较大，要从不同订单项的分布得到整个系列的控制曲线。

下面我们将分析问题并给出解决方案。

2 问题分析

2.1 定向条件

我们需要评估常见定向条件对于流量分布的影响。根据图 1统计目前平台、地域、频道和性别是定向最多的四个维度，同时每个订单项默认绑定到某个广告位，因此我们分析这五个维度，基于 5 月 25 日的数据，可以得出以下结论：

1. 除 041 广告位（暂停）的高峰出现在中午外，其他主要广告位流量分布相似，简洁起见，图 2a只列出了流量较大且有一定差异的广告位。
2. 不同平台（PC, iPhone, GPhone）的分布存在一定差异，PC 的分布和 041 广告位接近，可能和广告位流量在不同平台的占比有关。
3. 图 2c比较了北京、上海、河北、新疆四个省市，发现地域对流量分布的影响主要来自时差。
4. 图 2d比较了电视剧、动漫、资讯和汽车这四个比较有代表性的频道，由于受众不同，频道之间的流量分布存在一定差异，比如动漫频道的高峰较早，而汽车和资讯较晚。
5. 由于性别定向粒度较粗，对流量分布没有明显影响。

因此平滑消耗 1.0 只考虑广告位、平台、地域和频道四个定向，如果某个订单有其他定向条件，假设其不影响可触及流量按时间的分布，同时 1.0 版不考虑关键词定向。进一步，根据分析星期几对分布也有较大影响，所以我们使用上周同一天的流量生成当天的控制曲线，如果是国庆，小长假等使用最近一个节假日的数据。

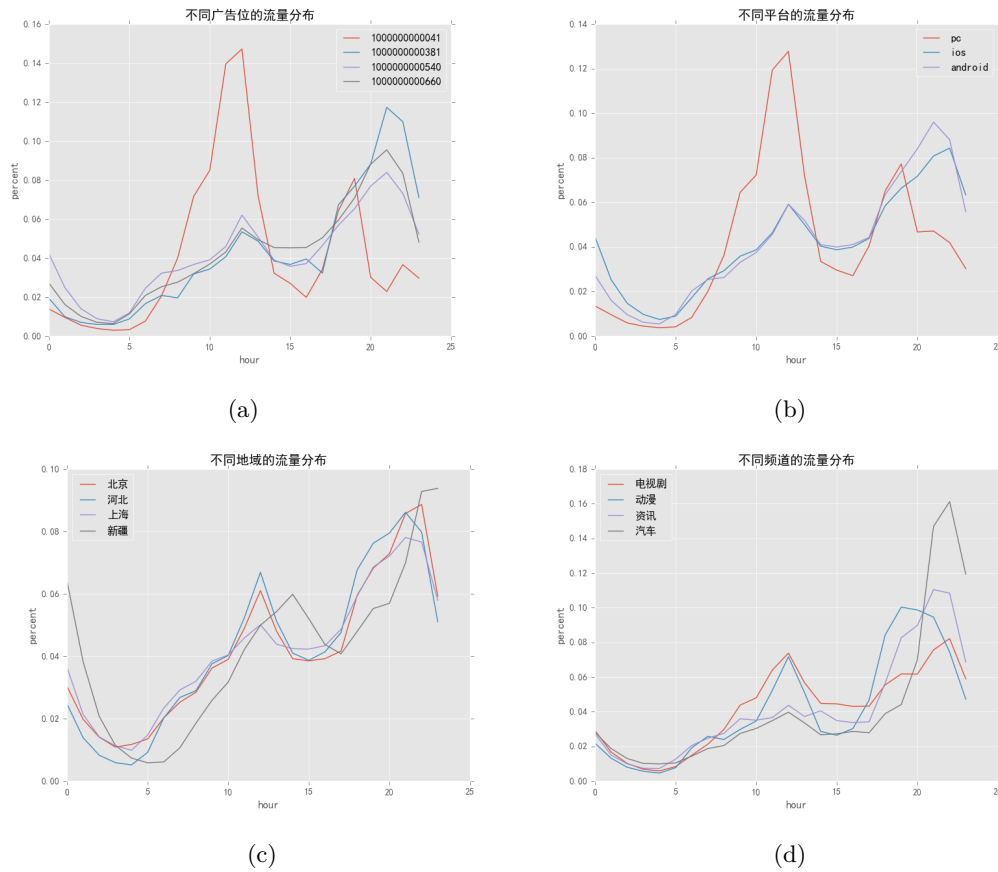


图 2: 单一定向条件对流量分布的影响

2.2 广告系列的控制曲线

奇效广告的天级预算设置在广告系列，一个广告系列可能包含多个订单项，我们统计了广告系列下的订单个数。由于 99% 的系列下的订单少于 20 个，所以图 3 只给出了包含个数小于 20 的广告系列分布，其中 67% (约 10 万) 的广告系列只有一个订单项。而一个系列的不同订单项可能对应不同的流量分布，如图 4 所示，同一广告系列下订单的流量曲线可能不同。

为了得到整个广告系列的控制曲线，我们需要将不同订单的分布进行组合。**建议 1.0 版本直接采用定向流量最大的订单项的分布作为整个系列的流量曲线。**控制曲线 $c(t)$ 由 24 个数插值确定 (可以直接线性插值, 将小时的流量增长平均到该小时的每分钟), $c(t)$ 代表 t 时刻结束时应该消耗的比例。由于某些系列由投放排期, 如设置为 18:30 到 20:29 投放, 则 $c(18.5h) = 0, c(20.5h) = 1$, 中间的值由对应流量分布插值决定, 由于排期越短我们预估流量偏差影响越大, 所以可能需要针对短排期系列增大加速因子, 如取 $\alpha = 0.2$ 。

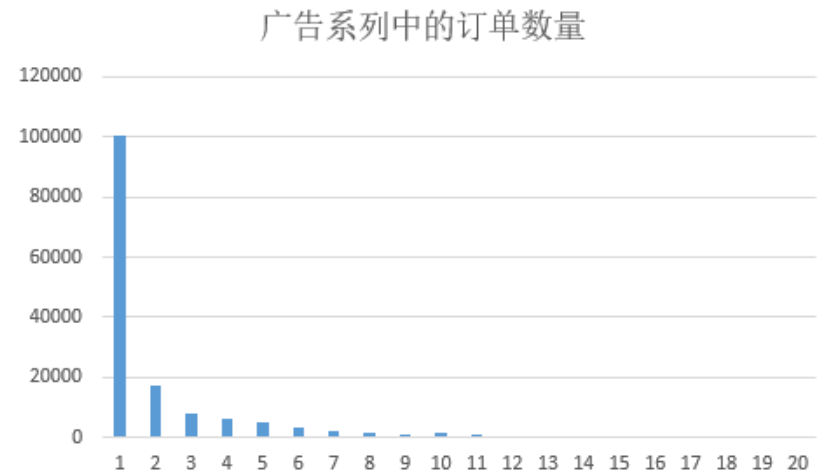


图 3: 包含订单项数的广告系列分布, 如有 10 万广告系列只有一个订单项

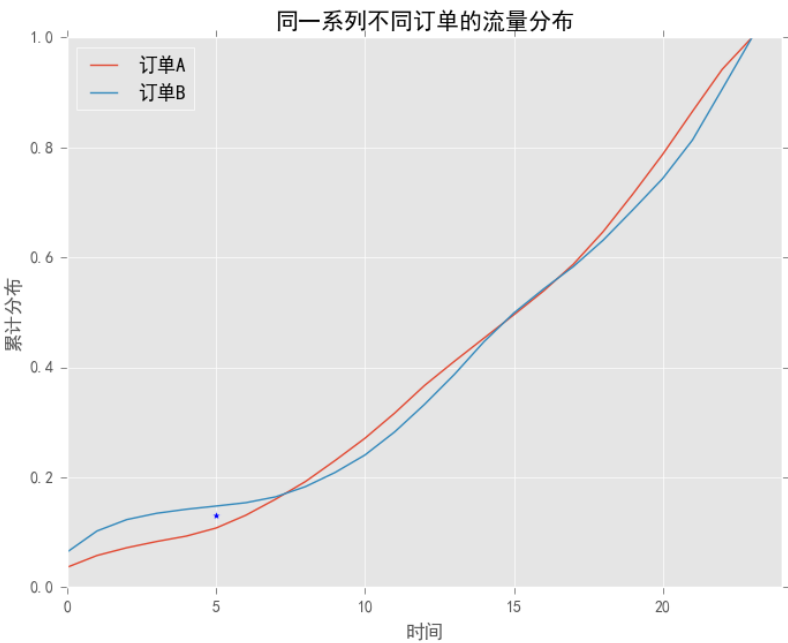


图 4: 同一系列两个订单项的流量分布

3 实现方案

3.1 基本架构

目前奇效已有的 预算控制 对每个系列计算概率 p , 用于控制其下的订单项进入 ranking 的概率。平滑消耗依然获取每个系列的 Ptr, 算法服务负责实时更新 redis 中每个订单项的 Ptr, 引擎的 impression handler 从 redis 取出即可。

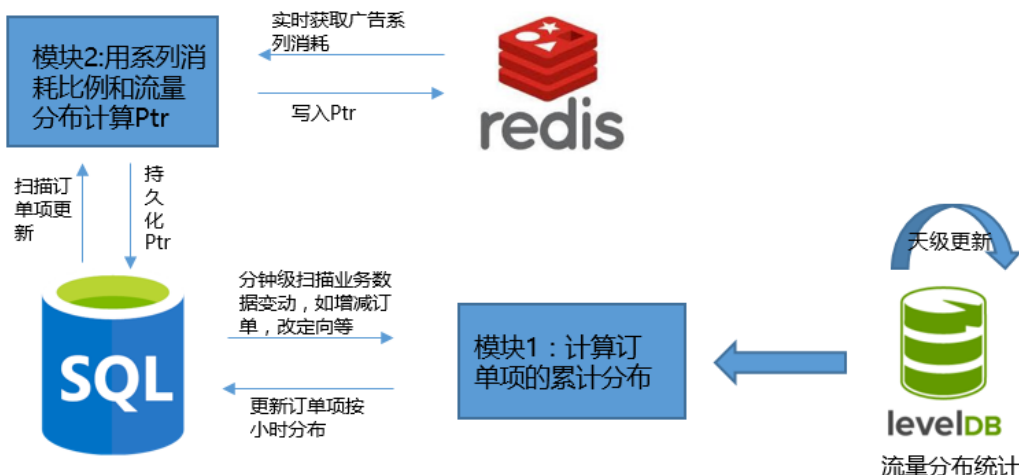


图 5: 基本架构

基本架构如图 5, 由于引擎 redis 中实时记录了每个系列的消耗, 所以在算法机器部署一个 redis slave 用于读取消耗数据。我们每天统计不同粒度下的按小时流量分布, 模块 1 用上周数据, 根据订单项定向条件计算其累积分布和总曝光数 (25 维向量), 将结果存入 mysql。由于广告主可能增减广告, 修改定向条件, 模块 1 需要每隔一段时间扫描业务数据库改动, 更新对应广告系列的分布。模块 2 每隔 1 分钟从 mysql 更新广告系列的控制曲线和预算, 从 redis slave 读取广告系列的消耗, 计算 Ptr, 写入引擎给定的 redis master 中。我们将订单项 Ptr 的 TTL 设置为 30 分钟, 如果由于服务异常, 超过 30 分钟未更新, key 将自动过期, 回退到老版本的预算控制。(1)的具体参数先定为 $r_t = 10\%$, $\alpha = 0.1$, 初始的 $p_i(0) = 1.0$, 每天 22 点后取消平滑消耗控制。初始概率等于 1, 可以兼容不限预算的情况。同时引擎还需要有对平滑消耗的保底策略, 即若平滑筛选后的订单集合为空时, 随机挑 N 个订单作为候选。

3.2 计算累积分布

平滑消耗中重要的一步是根据订单项的定向条件快速计算其累积分布。如果按小时聚合, 则每个订单项对应一个 24 维向量。由于 1.0 版只考虑广告位、平台、地域和频道四个定向条件, 因此我们需要根据定向条件查询不同粒度的累积分布和曝光数, 借助 levelDB 可以简单高效的实现 (当然其他 noSQL DB 亦可)。levelDB 是 Google 开源的硬盘化 Key-value 存储方案, 它的特点是高性能、易操作, 类似的工具还有 Facebook 开源的 RocksDB。我们在硬盘上新建一个 levelDB, 然后每天以日期 + 广告位 + 平台 + 地域 + 频道作为 key, 24 小时分布和曝光数 (25 维向量) 作为 value, 更新到 levelDB。同时模块 1 扫描业务数据库, 将订单项的定向条件做笛卡儿积, 从 levelDB 中查询不同粒度的分布和曝光, 按照曝光加权聚合。如订单 A 定向 540 广告位, iPhone, GPhone, 北京, 上海, 电视剧, 电影, 只需要遍历 $\{540\} \times \{\text{iPhone, GPhone}\} \times \{\text{北京, 上海}\} \times \{\text{电视剧, 电影}\}$ 组合的 key, 即可得到累积分布。

4 效果评估

常用的评估指标有：

- 广告系列在投时长：可以用广告系列消耗 95% 预算所需时间作为其在投时长；
- 广告的每单位消耗带来的 uv：订单项的 $uv/charge$ 越大说明其触达更多不同用户；
- 超投率：由于预算消耗更平滑，超投率应该有所降低；
- 预算未消耗量：平滑消耗可能减少预算的消耗，所以需要监控广告系列预算消耗不完的比例。

由于平滑消耗会对订单整天的曝光产生影响，所以不能采用通常的随机分割流量的方法 A/B 测试。我们使用对于一个系列切换开关的方式进行测试，具体方式见表 1。

	Sun	Mon	Tue	Wed	Thu	Fri	Sat
第一周	Off	On	Off	On	Off	On	Off
第二周	On	Off	On	Off	On	Off	On
第三周	Off	On	Off	On	Off	On	Of

表 1: 分天测试方案

参考文献

- [1] Agarwal, Deepak, et al., *Budget pacing for targeted online advertisements at linkedin.*, Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2014.