CT行业矩阵

1. 矩阵思路及设计

背景：

从2011/11月持续的数据统计中，发现CT广告行业和流量行业pair不同时，业务数据有较大差异；

从2011/11月CT上线后的bad case评估中，可以看到流量问题（如低俗流量，VC流量）和匹配问题（部分匹配或转义匹配）通常会伴随广告行业和流量行业的不匹配。引入行业矩阵，可以校验触发的广告推广组（unit）到页面（page）的匹配程度，对不适合展现的广告行业到流量行业对进行打压。

思路实现：

对每一个触发的unit，校验其广告行业与流量行业pair，如果命中矩阵，则按矩阵数据设定，用更高的截断阈值过滤（支持final score/pmr）；这样可以精准打压不合适展现的广告行业与流量行业pair。

1. 矩阵实施情况

* 全流量策略行业影响情况（实验数据，最后一列为人工diff评估的diff query内bad case比例变化）

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 展现 | 点击 | 收入 | CTR2 | CPM2 | ACP | JRATE | Bad Case(Diff) |
| B2B | 建筑 | -33.61% | -32.46% | - | 1.73% | - | - | 29.12% | -12.21% |
| 机械 | -52.66% | -38.95% | - | 28.96% | - | - | 39.66% | -12.76% |
| 节能环保 | -63.80% | -52.26% | -49.59% | 31.85% | 39.24% | 5.60% | 41.24% | - |
| 广告及包装 | -43.65% | -40.74% | -41.50% | 5.17% | 3.81% | -1.29% | 26.76% | -2.34% |
| 能源 | -61.63% | -34.94% | -34.61% | 69.55% | 70.40% | 0.50% | 17.30% | - |
| 纺织辅料 | -54.10% | -32.04% | -25.96% | 48.06% | 61.33% | 8.96% | 79.35% | - |
| 化工及材料 | -44.11% | -42.04% | -38.15% | 3.70% | 10.67% | 6.72% | 35.47% | -4.21% |
| 电子电工 | -53.68% | -40.90% | -37.21% | 27.59% | 35.57% | 6.25% | 22.37% | -6.37% |
| 安全安保 | -39.98% | -34.05% | -32.77% | 9.88% | 12.00% | 1.93% | 11.34% | - |
| 非B2B | 医疗健康 | -21.67% | -22.05% | -21.06% | -0.48% | 0.78% | 1.27% | 18.15% | -3.62% |
| 教育工作 | -9.88% | -8.40% | -9.62% | 1.64% | 0.29% | -1.33% | 4.31% | -3.57% |
| 交通运输 | -10.73% | -13.15% | -16.88% | -2.71% | -6.89% | -4.29% | 5.39% | -18.57% |
| 家居用品 | -36.21% | -28.50% | -25.57% | 12.08% | 16.68% | 4.11% | 27.61% | -5.85% |
| 低俗流量CT Total | | -1.78% | -1.63% | -2.60% | 0.15% | -0.83% | -0.98% | 1.89% | -62.96% |
| 低俗标记（冰峰）CT Total | | -5.15% | -1.14% | -5.52% | 4.23% | -0.39% | -4.43% | 2.65% | - |
| CT Total | | -20.26% | -16.56% | -24.27% | 4.66% | -5.02% | -9.24% | 19.86% | -26.43% |
| 反转实验 | | 25.41% | 19.84% | 32.05% | -4.45% | 5.29% | 10.18% | -16.57% | - |

* **未过滤bad case的原因分析**

**第一类，kate划分错误或未标注，约55%**

如抽样case 76，广告为招商加盟，广告kate未识别，匹配低俗流量的bad case；

**第二类，行业特性决定（即豁免pair中多数是good case的同时难以避免的bad case），约5%**

如抽样case 138，广告为安全安保，流量为建筑装修，均识别正确，豁免合理，但实际是安全门窗广告匹配建筑技术网页

**第三类，广告主的宽泛/人群提词达到full match，约40%**

如抽样case 141，广告为办公家具，流量为电脑设备，均识别正确，买词”复印机”，因full match豁免

**解决思路**

1. 第一类可以通过提升kate的准确召回（包括广告行业kate及流量行业kate两方面）解决；
2. 第二类可以考虑review策略pair，去除已设定的豁免pair，可能会误杀good case，有误杀率较高的风险（这块在策略制定时豁免的pair都是相对good case较多的pair）；
3. 第三类行业矩阵策略较难解决，虽然可以对非豁免行业pair设定强行过滤full match，但是较大损失召回

**参考数据**

1. QA对行业矩阵的性能测试，只保留完全一致广告行业到流量行业匹配的case豁免，其他均做full match截断的行业矩阵最大化极端情况下，bs会损失~75%的CT广告返回
2. 经验数据full match的good case率高于全局

* **误杀good case的原因分析（根据第一轮行业矩阵总体diff评估，总体在diff内占比约~6%，pv占比~0.8%）**

**第一类，策略特征错误，约占good case误杀65%**

原本豁免矩阵的pair，因广告行业或流量行业划分有误，或者ado标记错误而被过滤

**第二类，策略设定问题，约占good case误杀35%**

某些特定的pair，在特定条件下是good case，但是基于该pair的bad case量较大，未设定豁免导致误杀；或者原本是展现在图文广告位的good case文本广告，因同一广告位的bad case过滤，一并被图片广告取代导致good case被一并误杀

1. 后续计划执行情况

pfs-kate特征调研结论：

当前pfs-kate特征，一级准确率85%左右，召回率90%左右；

afs-kate特征调研结论：

当前afs-kate特征，使用广告主提词bidwords作为计算源，一级准确率85%左右，召回率98%以上

采用创意核心词+同landing page聚合（langding page截断?后参数）的方法，一级准确率90%以上，召回率75%左右

后期应用考虑分化为描述广告主投放意图的afs-bidwords-kate及描述广告主所属行业的afs-idea-kate

CT与QT效果比较差，且可以考虑使用矩阵改进的行业已筛选：

**教育及工作、交通运输、医疗健康**、游戏、商务服务、金融服务、旅游、**电脑办公及电脑软件**、招商加盟(加粗行业已有策略)