

# 大数据下的广告排序技术 及实践

蒋龙

2013年11月16日

# 提纲

- 广告排序的核心问题
  - 点击率预估
- 基于机器学习的点击率预估
  - 数据，特征，模型，评测
  - 大数据下的特征处理和模型训练
- 深入讨论



# 搜索推广

Query: 雪纺连衣裙

s.taobao.com/search?q=%D1%A9%B7%C4+%C1%AC%D2%C2%C8%B9&commend=all&ssid=s5-e&search\_type=item&sourceId=tb.index&initiative

0-algo 5-生活服务 a-技术新闻 iphone 技术 淘宝网内 新浪微博 淘宝网-淘!我喜欢 TechWeb.com.cn ... W806手机内部购机... 资产管理系統

好店推荐: 黄钻买家最爱买 回头客最多 服务态度最好 发货速度最快 描述评分最高

袖长: 短袖 无袖/背心裙 长袖 吊带裙 五分袖/中袖 七分袖 +多选

图案: 纯色 花色 圆点 碎花 条纹 卡通 +多选

女装: 连衣裙 (519.2万) 雪纺衫 (6.7万) 长袖连衣裙 (55.8万) 背心/吊带衫 (6038)  
大码服装 (7.5万) 半身裙 (9052)

童装/童鞋/亲子装: 儿童裙子 (7.0万)

你是不是想找: 大码雪纺连衣裙 碎花雪纺连衣裙 蕾丝雪纺连衣裙 雪纺连衣裙秋冬 vvv雪纺连衣裙 雪纺衫 牛仔连衣裙

所有宝贝 人气 天猫 二手 逛街  启用搜索定制 | 设置 1/100 < >

雪纺 连衣裙 确定  海外商品  货到付款  消费者保障  7天退换  正品保障  旺旺在线  亲, 合并同款换位置啦

默认 销量 ↓ 信用 ↓ 价格 ↓ 总价 ↓  -  所在地  款式 店铺 列表 大图

在逛街中找到“雪纺 连衣裙”相关宝贝: 应季新品 特价商品

您好 rainbowgbh, 点击查看“雪纺 连衣裙”相关宝贝在收藏过店铺、购买过店铺、免邮费、同城内搜索的结果。 立即查看

凤色蝶旗舰店 限时折扣 7折包邮

红纱雀 淘宝第一 ¥128 包邮 腰带

Caisaufen 采尚枫

仅限今日 159 包邮 元 ¥199.00 2012 女装 碎花 雪纺 连衣裙 夏季 新款 最近成交11685笔

em.taobao.com/item.htm?id=17363372220

# 定向推广

热卖单品

精品凉鞋

热卖女鞋

新款凉鞋

坡跟女凉鞋

休闲凉拖

女鞋凉鞋

美鞋

拖鞋

性感鱼嘴

舒适单鞋

更多热卖



疯抢!漫步云端 女士包臀收腰蕾丝花边连衣裙

¥ 598.00



Amovo纯可可脂巧克力8口味礼盒 零食 包邮

¥ 109.00



思加图 专柜正品 2012夏季新款 拖鞋凉拖凉鞋

¥ 238.00



0利润大牌高档退换包邮真丝连衣裙送彩票

¥ 800.00

您可能对这些宝贝感兴趣

缀好



缀好



缀好



缀好



# 广告展现流程

## 1. 候选广告选取

– 用户 -> Query <-> Bidwords <- 广告主

## 2. 广告排序

– 按ECPM(Effective cost per mille)排序

•  $ECPM = CTR * Bid$

– 广告平台受益最大，兼顾用户和广告主需求

## 3. 扣费

– GSP拍卖(Generalized Second Price Auction)机制

–  $CPC_i = (CTR_{i+1} * CPC_{i+1}) / CTR_i$



# 核心问题： 点击率

- 广告排序
  - $ECPM = CTR * Bid$
  - 排序时Bid已知，但CTR未知
- 扣费
  - $CPC_i = (CTR_{i+1} * CPC_{i+1}) / CTR_i$
  - 当前广告的扣费依赖当前及后一条广告的CTR
- 所以，计算每条广告的CTR是排序和扣费的核心
  - $P(\text{click} | \text{query}, \text{ad})$



# 直接估计

- 广告每次被展现后有两种可能的结果：点击和不点击；
  - 假设点击概率为 $p$ ，则不点击概率为 $1-p$
- 假设 $p$ 恒定，则广告在 $n$ 次展现中被点击的次数 $X$ 服从二项分布。被点击 $k$ 次的概率为

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} = b(k; n, p)$$
$$(k = 0, 1, \dots, n),$$

- 基于历史统计的点击率估计
  - 例如，给定query时，某广告被 $n$ 次展现 $k$ 次点击，根据最大似然估计
  - $ctr = k/n$



# 直接估计的不足

- 直接估计的不足
  - 数据稀疏
    - 1千万广告，1千万query->100万亿pair
    - 新query，新广告
  - 真实点击率低，需要大量展现数据才能得到可靠估计
    - An ad with a true CTR of 5% must be shown 1000 times before we are even 85% confident that our estimate is within 1% of the true CTR
    - 大量尝试性展现浪费流量价值
  - 点击率未必恒定
- 解决方案：利用广告和query的各种特征，通过机器学习模型来预测



# 基于机器学习的点击率预估

- 问题类型
  - 分类：点击or不点击，但需要点击的概率
  - 回归：直接拟合点击率
- 模型和算法应尽量简单，易于并行化
  - 海量数据
    - 训练数据，特征
  - 快速更新需求
- 常用方法
  - 逻辑回归
  - Boosting类，如GBRT



# 逻辑回归模型

- 二元逻辑回归(Logistic Regression, LR)模型

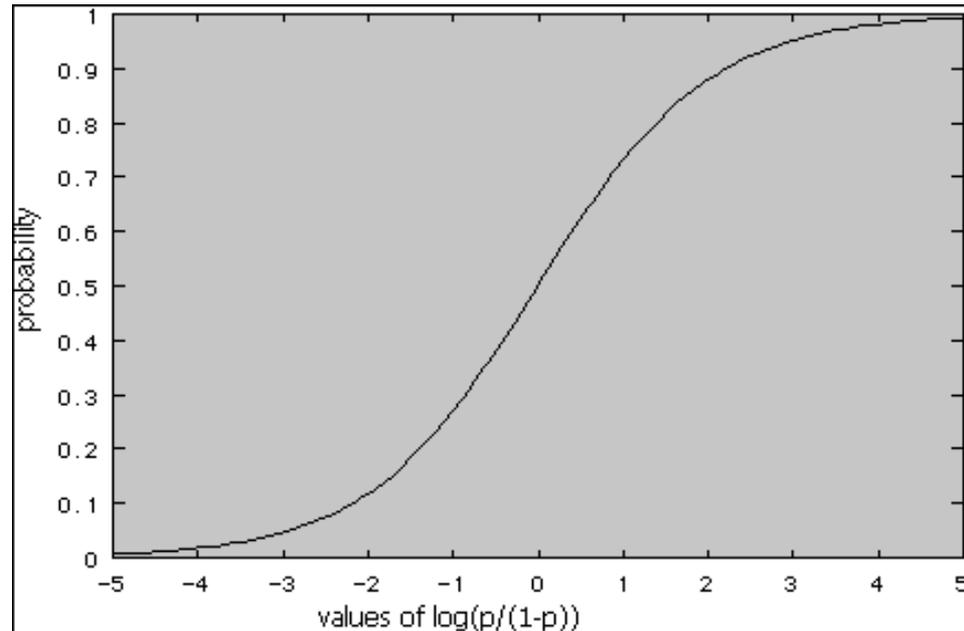
$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\sum\beta_i*x_i)}} \quad P(y=0|x) = \frac{e^{-(\beta_0+\sum\beta_i*x_i)}}{1+e^{-(\beta_0+\sum\beta_i*x_i)}}$$

- 几率(odds)

$$\frac{p(y=1)}{p(y=0)} = e^{(\beta_0+\sum\beta_i*x_i)}$$

- 对数形式

$$\ln\left(\frac{p(y=1)}{p(y=0)}\right) = \beta_0 + \sum\beta_i * x_i$$



# 基于LR的点击率预测

- 基于LR的点击模型

$$P(y=1|x) = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_i * x_i)}}$$

- Where,  $x$ 代表一个(query, ad)对应的特征向量,  $y$ 属于{1,0}分别代表点击和不点击,  $p(y=1|x)$ 就代表给定query下, 某ad的点击率



# 点击率预测特征

- 广告创意特征
  - 图片，标题文字，价格，销量
  - 推广商品所属类目，包含属性
  - 创意组，推广计划，广告主
- Query信息
  - 包含的Terms
  - Query分析：类目，属性
  - Query扩展：同义词，相似query
- 环境特征
  - 用户，时间



# 点击率预测特征

- 名义特征
  - 时间，创意ID等
- 点击反馈特征
  - 计算历史上包含该特征的(query, ad)的点击率
  - E.g., ad所属广告计划的历史点击率
- 组合特征
  - E.g., query与ad标题匹配的term个数



# 模型训练

- 训练/测试数据
  - 一定时间内的广告系统日志
    - (query, ad): click, pv-click
  - E.g., 1个月数据训练，接下来的1天数据测试

- 训练目标：最小化负对数似然

– 似然函数： $\prod_l P(Y^l | X^l, \beta)$

– 负对数似然：

$$l(\beta) = -\sum_l \log p(Y^l | X^l, \beta)$$

$$= -\sum_l [L_l^+ \log P(Y^l = 1 | X^l, \beta) + L_l^- \log P(Y^l = 0 | X^l, \beta)]$$



# 正则化

- 通过加上正则项，得到稀疏模型

$$\min_{\beta} - \sum_l [L_l^+ \log P(Y^l = 1 | X^l, \beta) + L_l^- \log P(Y^l = 0 | X^l, \beta)] + \lambda \sum_i |\beta_i|$$

- 通过正则项控制模型复杂度，避免over-fitting
- L1正则项能使大量的无效特征权重为0，起到特征选择作用



# 模型训练

- 参数估计算法
  - 梯度下降
  - 牛顿法
  - BFGS
- L-BFGS
  - 拟牛顿法的一种，利用有限内存近似BFGS算法
  - 利用历史值和梯度寻找当前方向(Two loop)
  - Line search确定步长



# 效果评估

- 离线自动评估
  - 在测试数据上，评测模型的AUC，MSE指标
- 离线人工评估
  - 通过随机选取的query集合来对比新旧版本的相关性效果
- 线上评估
  - 把模型部署到线上，分配一部分真实流量来评估模型效果(CTR, ECPM)



# AUC指标

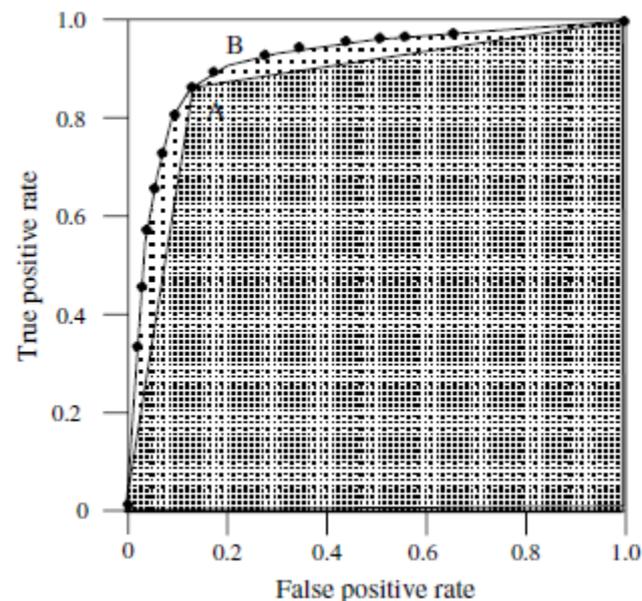
- Receiver Operating Characteristics (ROC) graph

		True class	
		p	n
Hypothesized class	Y	True Positives	False Positives
	N	False Negatives	True Negatives

fp rate =  $\frac{FP}{N}$       tp rate =  $\frac{TP}{P}$

precision =  $\frac{TP}{TP+FP}$       recall =  $\frac{TP}{P}$

accuracy =  $\frac{TP+TN}{P+N}$



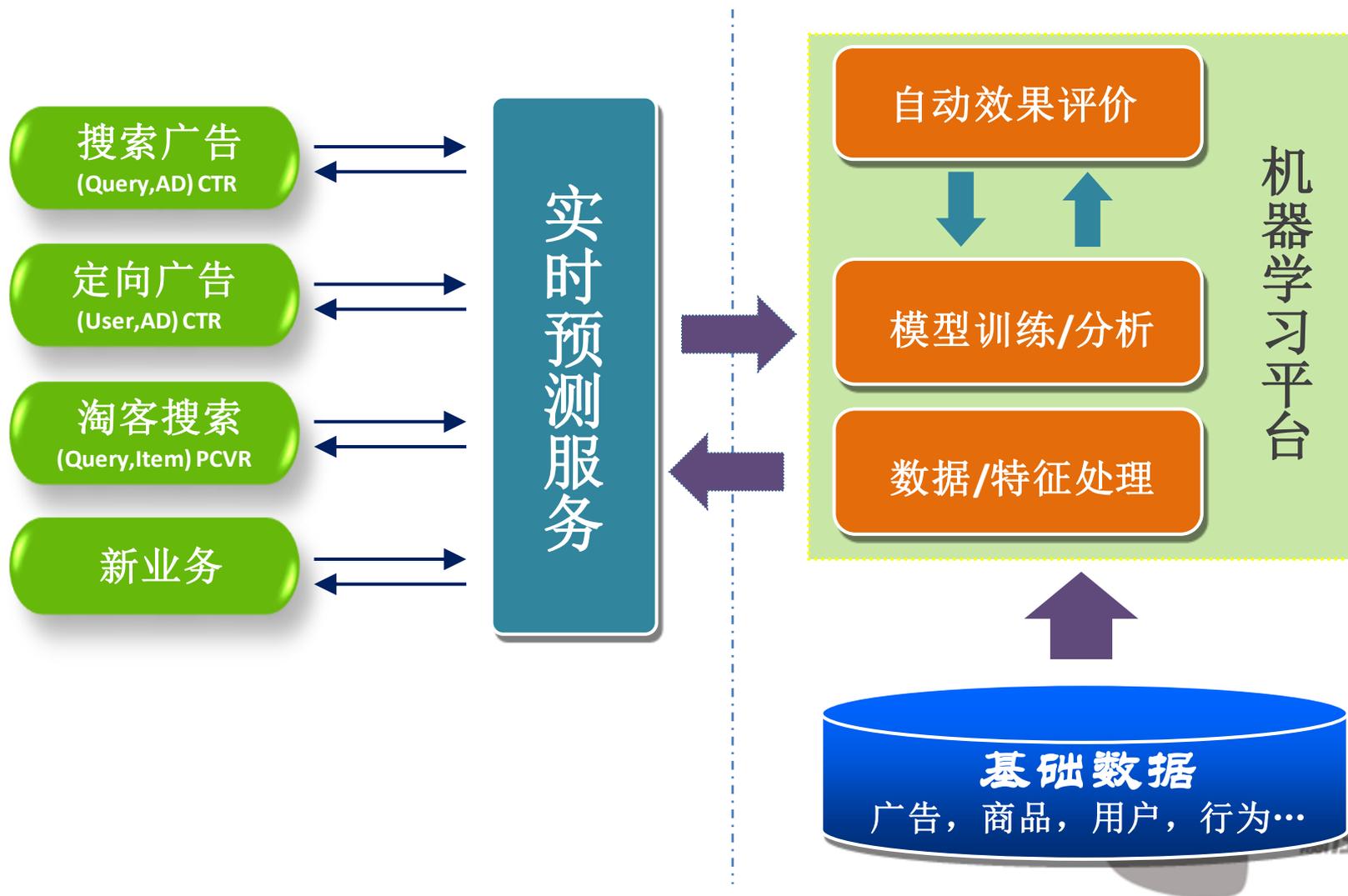
- Area under an ROC curve (AUC)
  - FPRate-TPRate曲线下方的面积
  - 衡量了排序的准确性
  - 越接近1模型越好，0.5是随机模型

# 大数据下的点击率预测

- 大数据挑战
  - 海量特征
  - 海量训练样例
  - 模型快速更新，多模型试验
- 解决方案
  - 数据/特征处理
    - 基于Hadoop的并行数据处理
  - 模型训练
    - 基于MPI的并行训练算法



# 大规模机器学习平台



# 大数据处理平台

- 大规模MapReduce+HDFS集群
  - 4000+节点，80PB+存储
  - 搜索广告点击率预估
    - 原始日志分析：10T+
    - 特征提取：吞吐50T
    - 训练数据：20T
- 基于MPI的机器学习算法平台
  - 500+台机器（单机24 CPUs, 96G内存）
  - 自动任务调度和监控系统，部署多种算法，如LR，PLSA，LDA，SVM，GBDT等
  - 搜索广告点击率预估
    - 十亿级特征，百亿级训练样例
    - 运算时间：~2小时



# MPI与Map-Reduce

- MPI(**M**essage **P**assing **I**nterface): 消息传递接口
  - 消息传递函数库的标准规范
  - 通过在进程间传递消息完成数据交换, 如Send, Recv
  - 程序员可以深入控制数据交互
- MPI VS. Map-Reduce模型
  - MPI: 适合逻辑复杂的迭代运算, 如机器学习算法
  - MR: 适合计算独立, 迭代少的任务





# 深入讨论-位置偏差

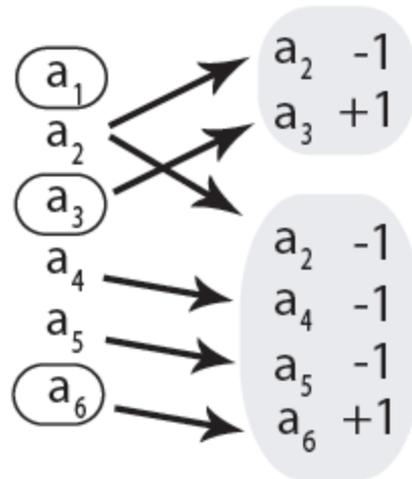
- 位置偏差(position bias)
  - 假设：不同位置上的广告被用户看到的概率不同，排位靠前的广告被看到的概率更大，导致其点击率“天然”更高一些
- 解决方案一：
  - (Cheng and Cantú-Paz, 2010)
    - 计算广告CTR时用不同排位上的平均ctr进行调整

$$\text{COEC} = \frac{\sum_{r=1}^R c_r}{\sum_{r=1}^R i_r * \text{CTR}_r}$$



# 深入讨论-位置偏差

- 解决方案二：
  - Online Learning from Click Data for Sponsored Search. (Ciaramita et al., 2008)
    - 只有当排位低的广告被点击而排位高的广告没有被点击时，才使用这些广告作为训练数据



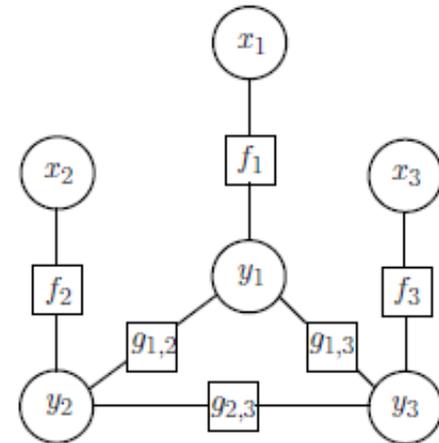
# 深入讨论-个性化

- Personalized Click Prediction in Sponsored Search.
  - Haibin Cheng, Erick Cantú-Paz. WSDM. 2010.
- 假设：query相同时，不同user对不同ad的点击率也相差较大
- 解决方案：加入user特征到LR模型里，预估  $p(c|q,a,u)$ 
  - Demographic特征
    - 如年龄，性别，婚姻状况，职业，兴趣等
  - User-specific特征
    - 如用户历史CTR，user组合特征(User-Ad, User-Query)

# 深入讨论-广告间相互影响

- Relational Click Prediction for Sponsored Search
  - Chenyan Xiong, Taifeng Wang, Wenkui Ding, Yidong Shen, Tie-Yan Liu. WSDM 2012.
- 假设：某条广告的点击率会受到同时展现的其他广告的影响
- 解决方案：同时预估所有广告的点击率，考虑广告间的相互影响
  - 广告间的相似度
  - Continuous CRF model with MLE

$$P(Y|X) = \frac{1}{Z(X)} \exp \left\{ \sum_i h(y_i, X; w) + \sum_{j>i} \beta g(y_i, y_j, X) \right\}$$



Thanks!

