DTCC

# 2016中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2016

数 据 定 义 未 来

SequeMedia 盛拓传媒　ITT168.com　ChinaUnix　ITPUB

# 在线学习在广告中的应用

王兴星

# 互联网广告简史

- 互联网简史：

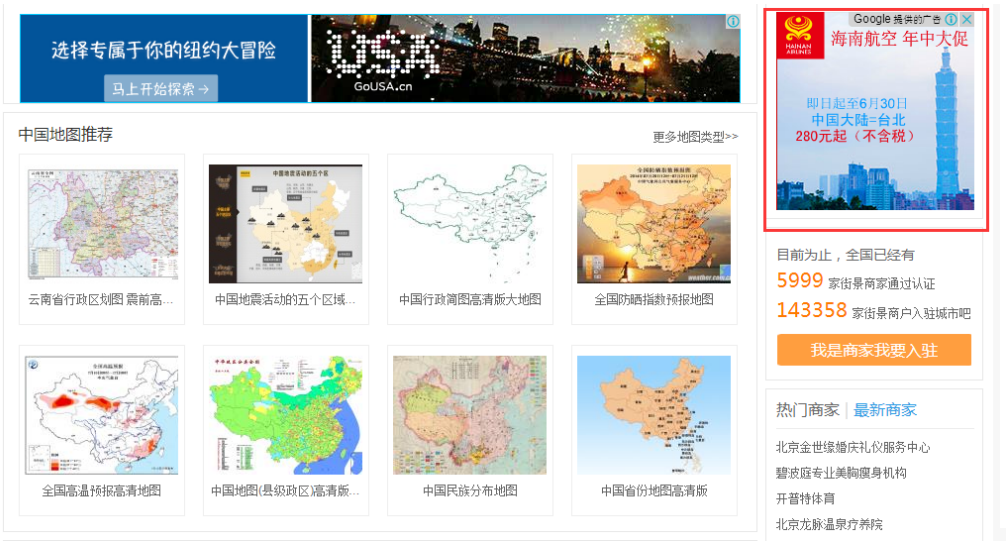
  门户-->搜索引擎-->垂直类网站-->无线；

- 互联网广告简史：

# 搜索广告


Bai du 百度
iphone6
百度一下
添加百度到桌面，搜索更便捷!
网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»
百度为您找到相关结果约43,300,000个
搜索工具
<京东>全新苹果6现货抢!你分期我免息!
推广链接
JD.COM 京东
网络：移动4G | 联通4G | 电信4G | 移动3G | 更多»
价格：0-5299 | 5300-5799 | 5800-5999 | 更多»
大家说：屏幕大 | 待机时间长 | 电池耐用 | 更多»
www.jd.com 2015-04
Apple Store苹果官网 - Apple官网在线商店(中国)
全新 iPhone 6 及 iPhone 6 Plus开售,享受免费送货,免息分期付款等服务(以网站为准).配备A8芯片,现有金色,银色,深空灰色.当止于大,效能非凡.
store.apple.com 2015-04 - 评价
iphone6 <中国电信网上营业厅> 春季钜惠!
<中国电信>iphone6, 79元套餐加增200元话费, 700流量, 前所未见, 空前直降!<中国电信>iphone6, 100%正品行货, 全国包邮, 详情咨询中国电信官方网站.
www.189.cn 2015-04 - 评价
苹果 iphone6 报价 参数 图片 评论 ZOL中关村在线
参考价格：￥4600-5288
模式：联通3G(WCDMA),电信3G(CDMA200 摄像：800万像素
主屏：4.7英寸 1334x750像素 系统：iOS 8.0
可选颜色：
参数 商家 图片 点评 论坛 二手
热门型号 摄像 参考价格
百度搜索 有V有保障
相关电器
展开
iphone 7
华为mate7
透明手机
iphone 6s
ipad mini 3
小米3s
window phone
红米note
其他人还搜
展开
三星
iphone 6c
荣耀6plus
iphone6 plus

# 展示广告


选择专属于你的纽约大冒险
马上开始探索 →
USA
GoUSA.cn
中国地图推荐
更多地图类型>>
云南省行政区划图 震前高...
中国地震活动的五个区域...
中国行政简图高清版大地图
全国防晒指数预报地图
全国高温预报高清地图
中国地图(县级政区)高清版...
中国民族分布地图
中国省份地图高清版


Google 提供的广告
海南航空 年中大促
即日起至6月30日
中国大陆=台北
280元起（不含税）
目前为止，全国已经有
5999 家街景商家通过认证
143358 家街景商户入驻城市吧
我是商家我要入驻
热门商家 | 最新商家
北京金世缘婚庆礼仪服务中心
碧波庭专业美胸瘦身机构
开普特体育
北京龙脉温泉疗养院


手创
电脑维修
报修热线：
13520353787
修
关闭
我们网站已正式开通！
欢迎各位新老客户来我公司参观指导工作，我公司具体的地址是：深圳市南山区桃源街道丽山路大学城创业园604室。
3dmax建模学习
3dmax学习班
编程基础知识
linux教程
北京地铁招聘
adwords
产品展示
欢迎您到访我们网站，目前主要的产品服务有， 如果对我们的产品感兴趣或者有任何的疑问，您可以直接给我们留言或直接与我们联系，我们将在收到您的信息后，将会第一时间及时与您联络。
地图
（下图中的红点是深圳市创领图像技术有限公司在的具体位置，地图可以拖动，双击放大）
深圳市创领图像技术有限公司
深圳市南山区桃源街道丽山路大学城创业园604室
西丽湖
桃源街道
大学城
塘朗
3dmax建模学习
3dmax学习班
编程基础知识
linux视频教程
电脑编程入门
3dmax学习
linux入门教程
linux教程
燕京啤酒代理商
北京旅游
电脑职业培训
游戏编程入门
自学编程
牛逼的站长
都在用它
防垃圾注册
防恶意灌水
防账户盗用
关闭

# 无线：信息流广告


推荐
热点
社会
娱乐
科技
2014.04.03星期四
今天
肯德基菜单大革新，15款产品新登场！4大星卖家你挺谁？
KFC
评论146 刚刚
KFC
谁是超强星卖家
由你说出来
夫妻争吵女人别轻易流泪
每天学点心理学
评论101 1分钟前
小贝儿子迷倒网友 盘点星二代正太帅哥

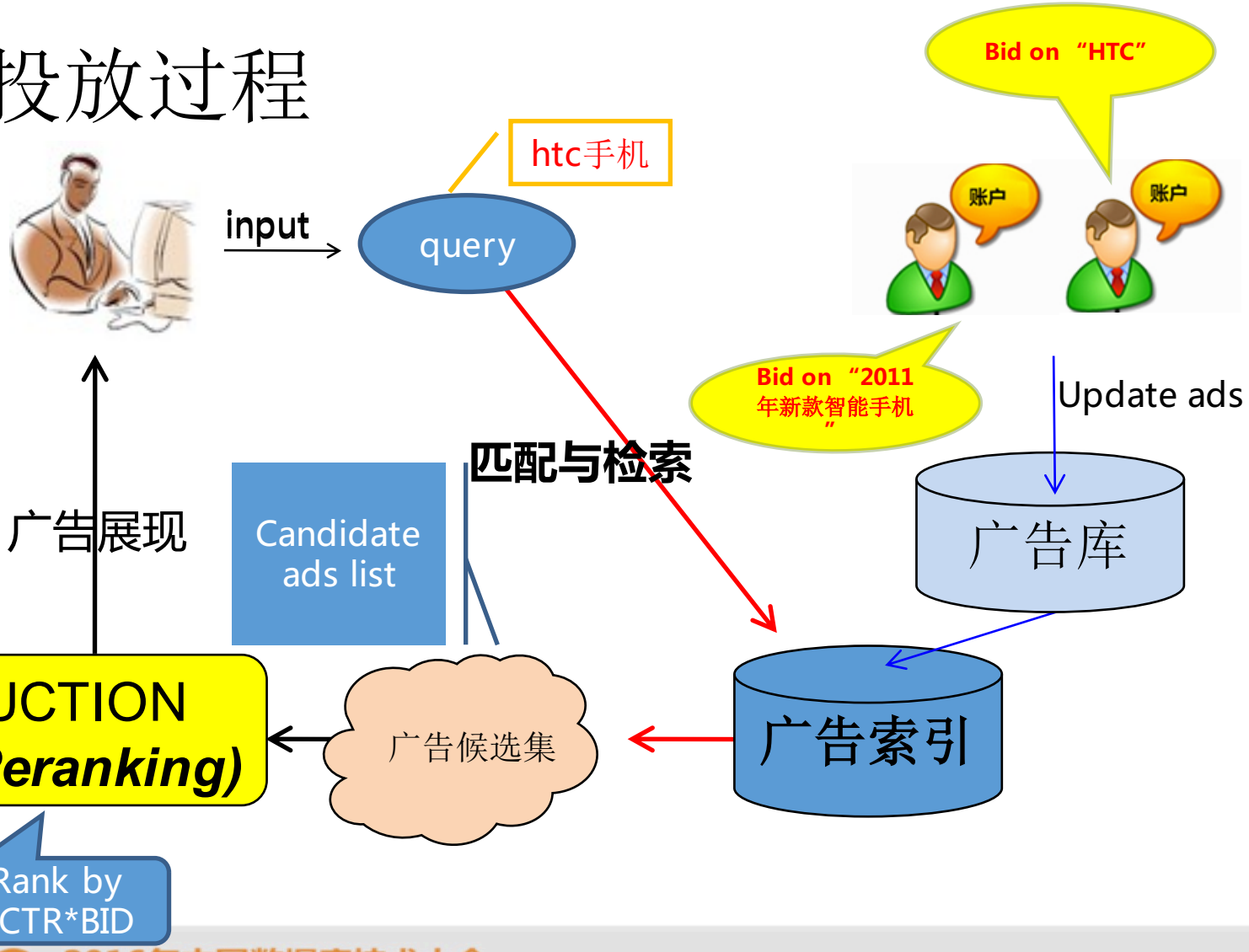

推荐
热点
街拍
电影
奇葩事
周迅
为新戏徐璐剪去长发，网友：我知道你很不适应，但是放心你好帅
田鼠叨叨叨 16评 2小时前
急用钱找宜人贷 无抵押 最快2小时放款！最高50万
立即下载
十分钟借十万
快至 2小时 到帐
999元春季好礼
「贷」你马上出发
凭信用卡
10分钟轻松借10万
推广
宜人贷借款 8.63万人下载
走路带风！金星素颜现身机场 大步流星
凤凰图片 12评 3小时前

# 问题规模

- 规模

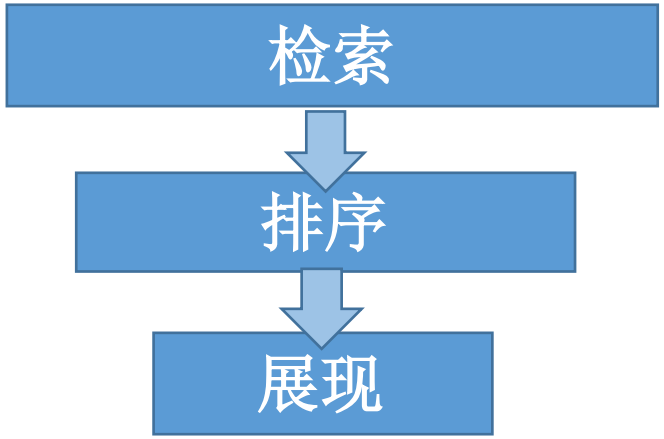
| | 搜索广告 | 展示广告 |
|---|---|---|
| 建模对象 | <Query, Ad> | <Ad, User, Context> |
| 索引规模 | 十亿 | 百万~千万 |
| 特征规模 | 亿~千亿 | 亿~千亿 |

# 广告投放过程


Bid on “HTC”
htc手机
input
query
账户
账户
Bid on “2011年新款智能手机”
Update ads
匹配与检索
广告展现
Candidate ads list
广告库
AUCTION (Reranking)
广告候选集
广告索引
Rank by eCTR*BID

# 相关技术

- 漏斗模型：




- 核心技术：

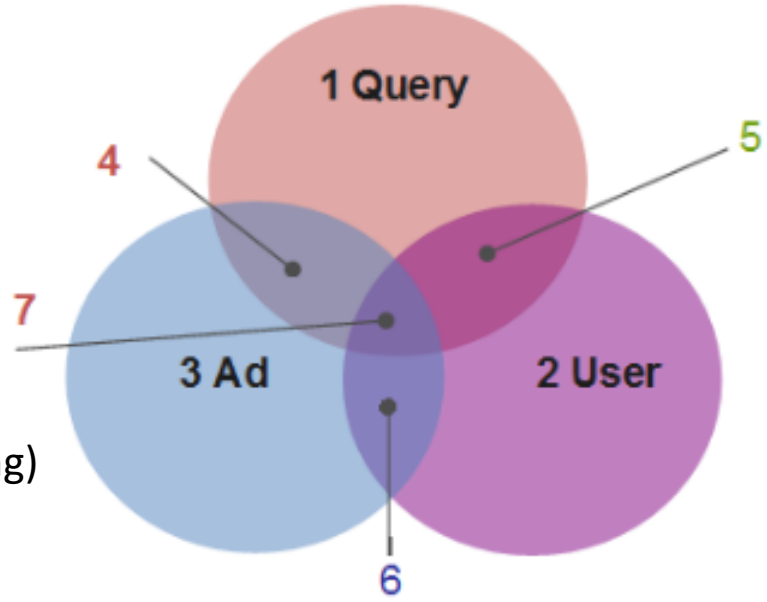
  点击预估：**Alive (Ad Living CTR Predict System)**

  信息检索：**L2M (Online Learning to Match)**

  物料技术；

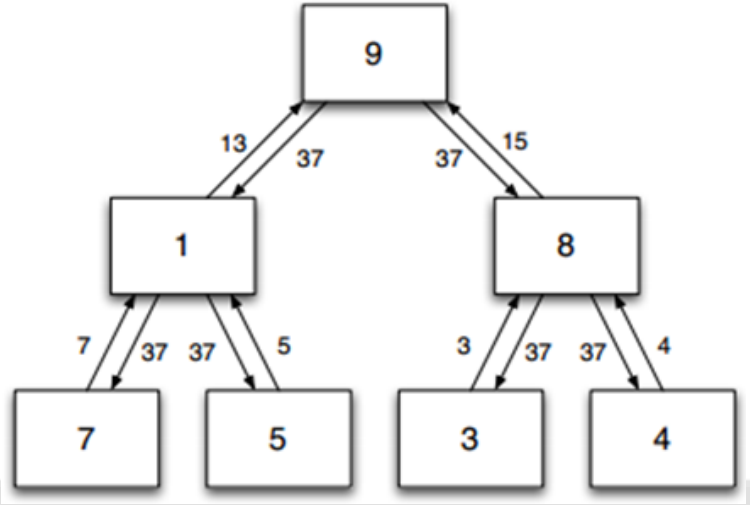
# CTR预估

- 常见技术方案：
  - 单特征+非线性模型：
    - 特征工作简单；
    - 计算资源消耗大；
    - 线上预估时间较长；
  - 单特征+组合特征+线性模型；
    - 人工/算法进行组合特征；
    - 传统方法计算资源中等；(Batch Learning)
    - 线上预估时间较短；

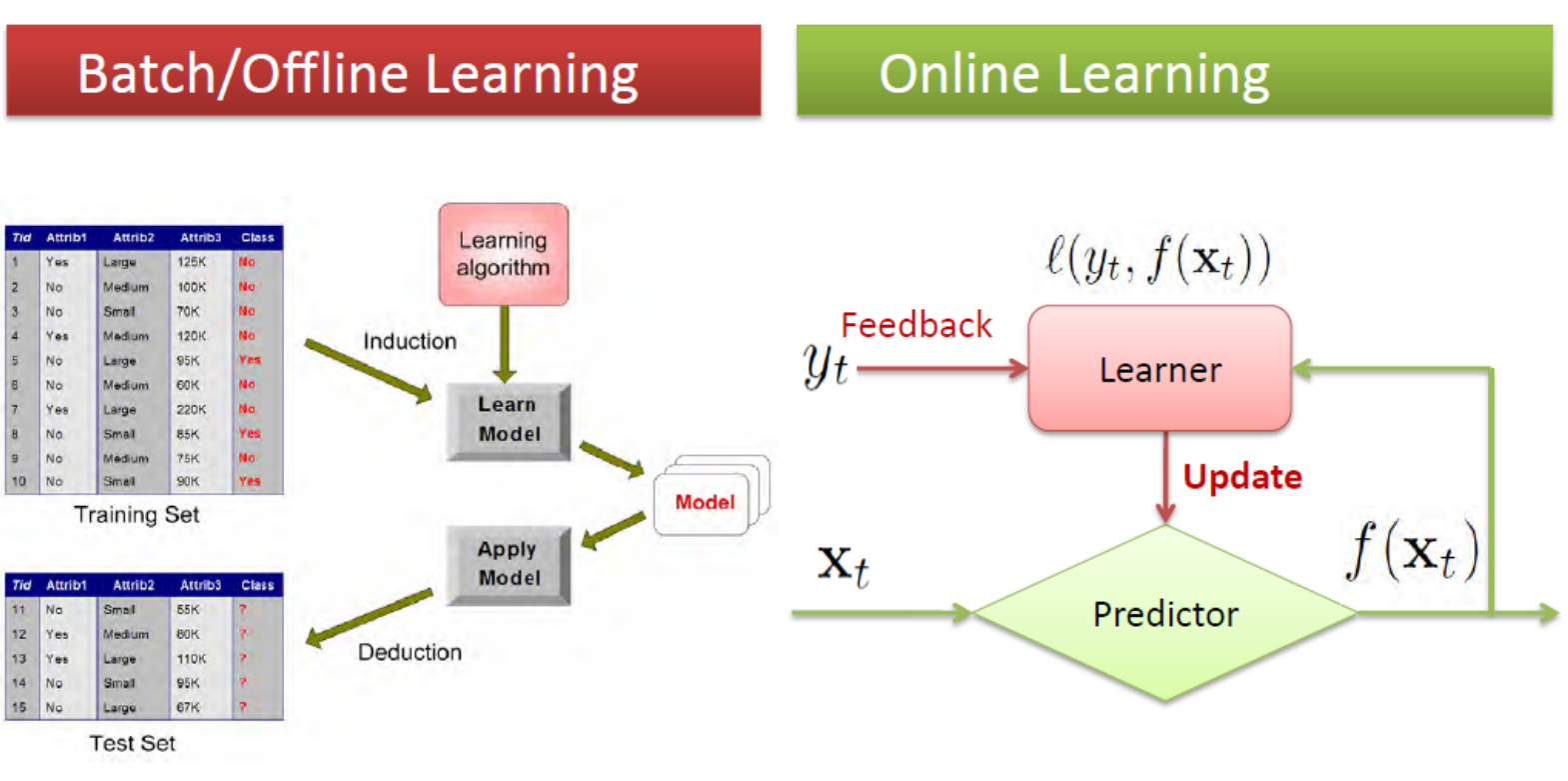


# 批量学习：Batch Learning

- 常见数据量：
  - 百亿级样本/每天，每次使用30~90天；
  - 亿~百亿级别特征；
  - 几十台机器，百轮左右，训练10个小时左右；



# Batch Learning vs Online Learning



# 理论依据

- $f^*$ 为全局最优解：

$$f^*(.) = \arg\min_{f \in H} l(y_t, f_t(x_t))$$

- 在线算法的Regret：

$$regret = \frac{1}{T}\sum_{t=1}^{T}[l(y_t, f_t(x_t)) - l(y_t, f^*(x_t))]$$

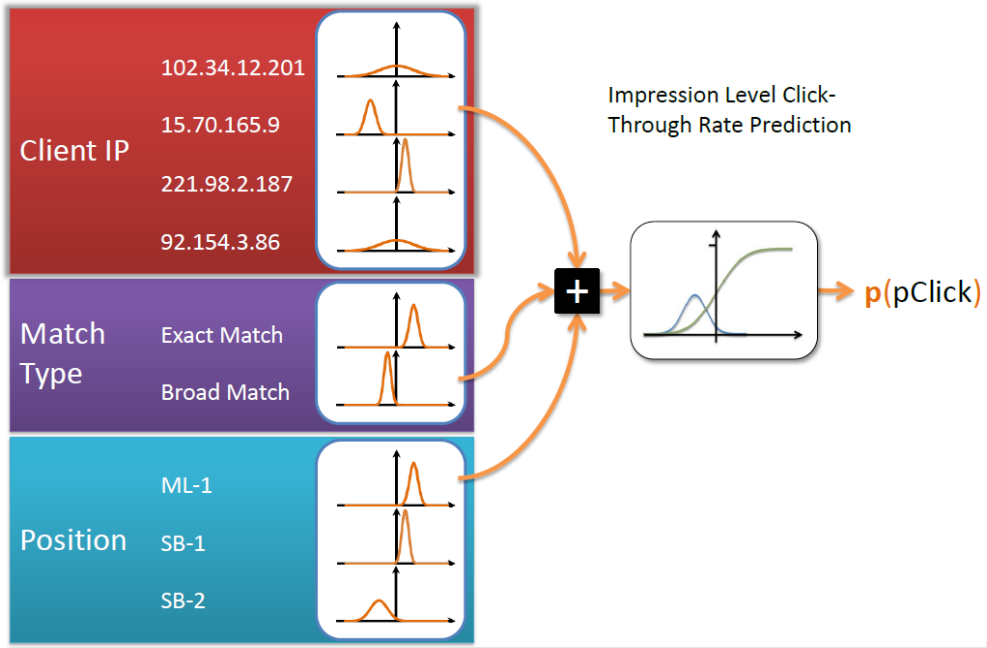
- 目标：希望Regret尽可能小
  - 保证在线算法的效果，与看到整个样本得到的算法的效果，尽可能接近。

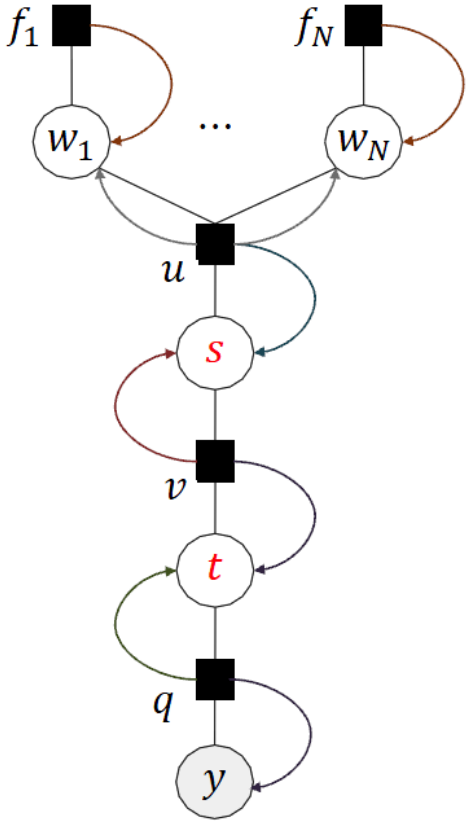
# 在线稀疏模型

- 在线稀疏模型：
  - Truncated  Weight；
  - Truncated Gradient(Langford, 2009)；
  - FOBOS(Duchi,2000)；
  - RDA(Xiao,2010)；
  - FTRL(McMahan,2010)；

# 在线模型：贝叶斯派

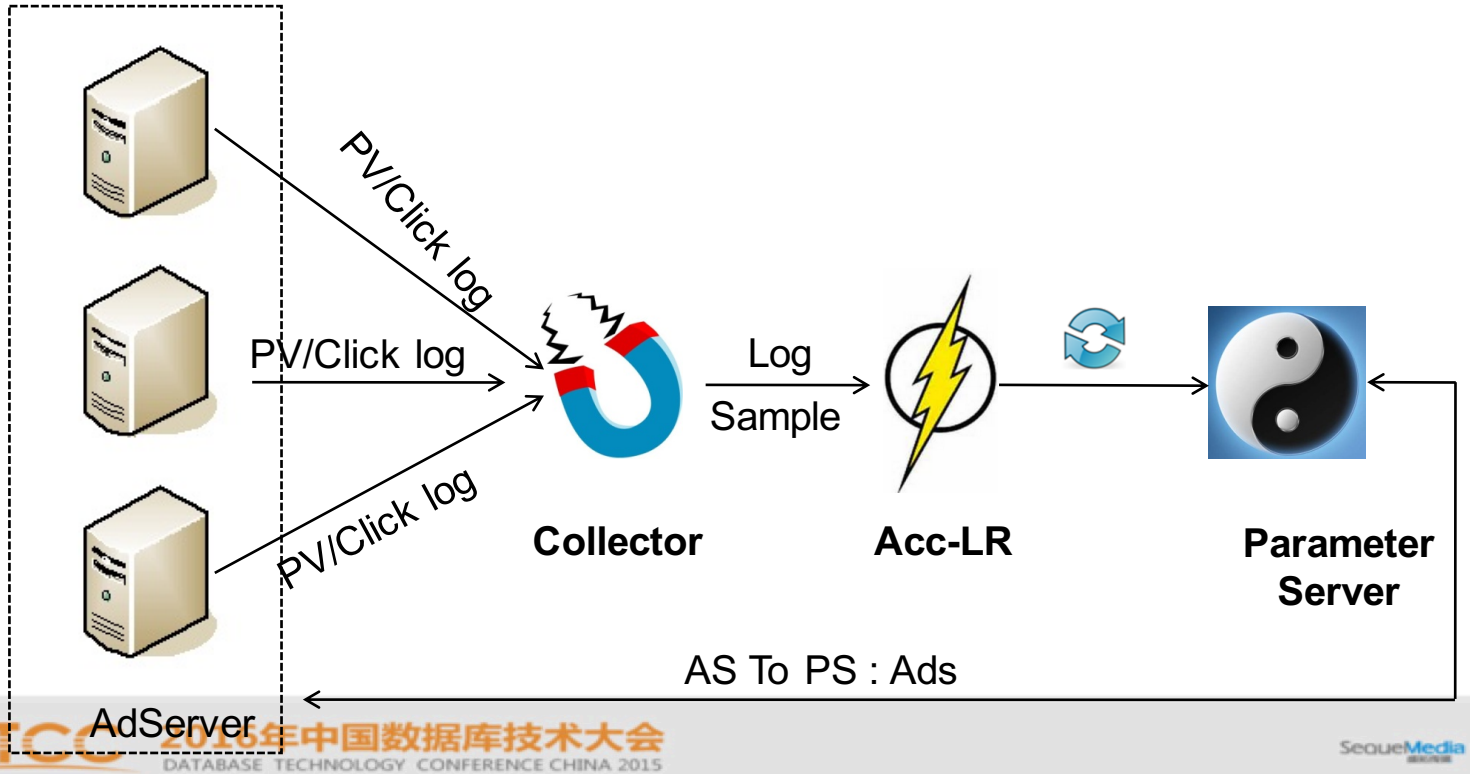
- AdPredictor:



# 在线模型框架

- 整体框架：


PV/Click log
PV/Click log
PV/Click log
Log
Sample
Collector
Acc-LR
Parameter
Server
AS To PS : Ads
AdServer

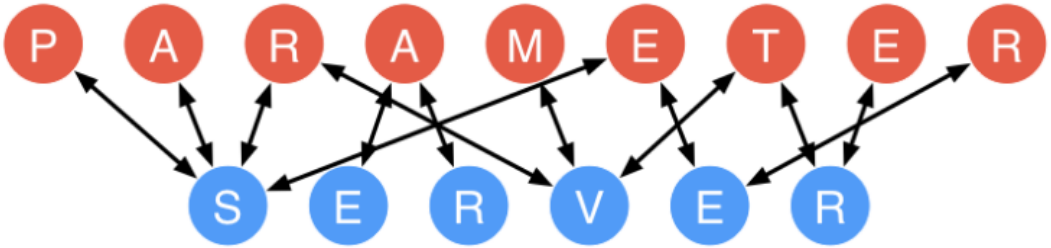
# 大规模学习解决方案

- 采样：正负样本失衡+多线程
- 同步方案：MiniBatch+AllReduce

**Algorithm 2** ParallelSGD($\{c^1, \ldots c^m\}, T, \eta, w_0, k$)

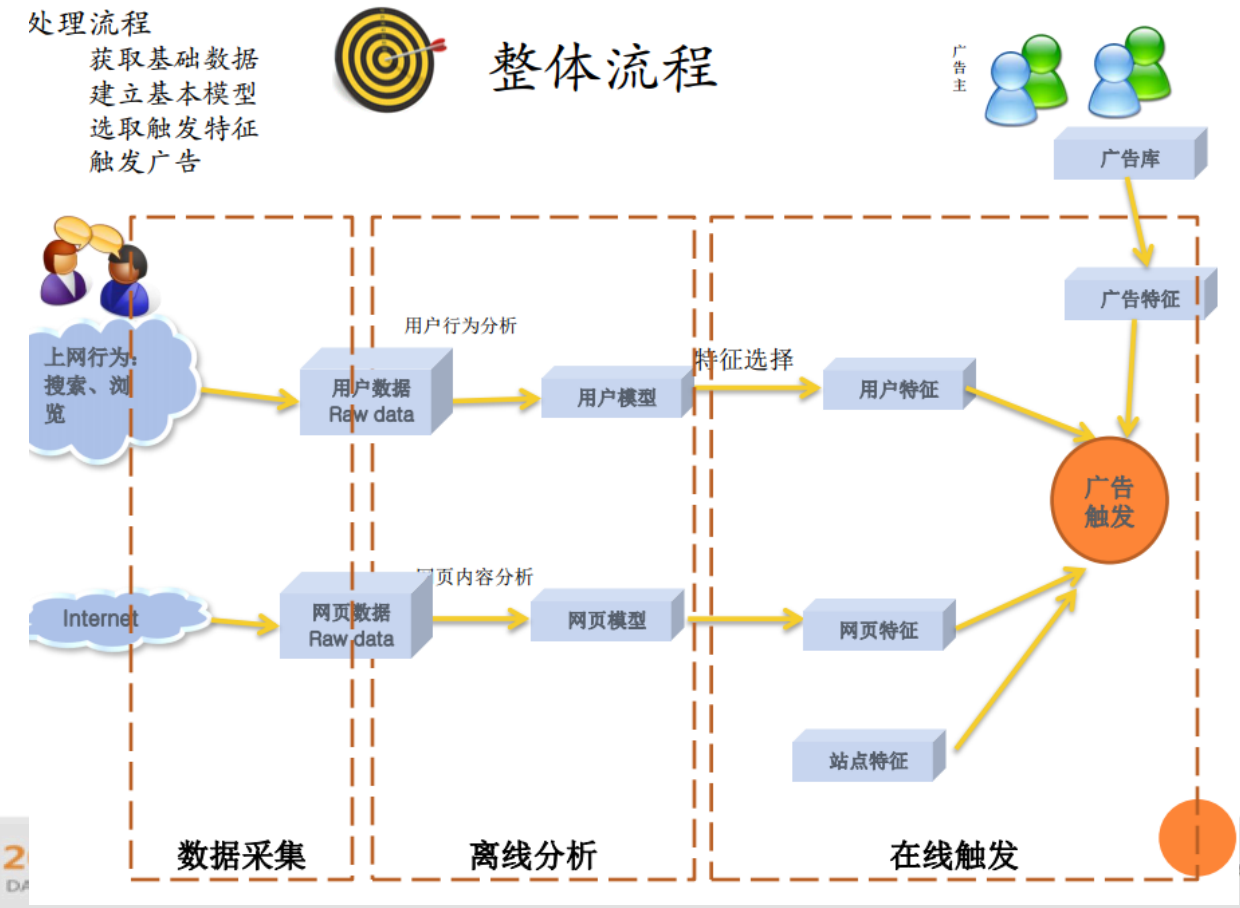
**for all** $i \in \{1, \ldots k\}$ **parallel do**
  $v_i = \mathrm{SGD}(\{c^1, \ldots c^m\}, T, \eta, w_0)$ on client
**end for**
Aggregate from all computers $v = \frac{1}{k} \sum_{i=1}^{k} v_i$ and **return** $v$

- 异步方案：MiniBatch+PS



# L2M： Online Learning To Match

- 流程：



# 常见触发方法

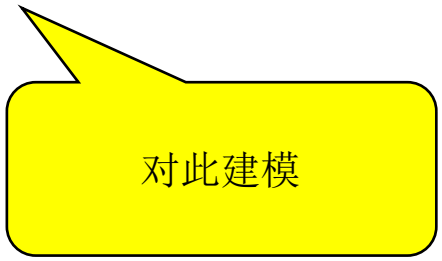
- 触发方法分类：
  - 基于用户：
    - 历史Query、兴趣标签；
  - 基于上下文：
    - 当前页面Title、历史浏览页面；
  - 基于站点：
    - 站点类别；
  - 基于重定向：
    - 历史点击；
  - 相似用户：
    - Look-alike；

# 无线触发遇到的问题

- PC-->无线，传统触发的问题？
  - 出发点：
    - 信息检索，非CPM；
  - 数据覆盖：
    - 用户数据：覆盖率低；
    - 网页数据：无商业性（小说、新闻）；

# L2M

- 出发点：
  - 触发本质：IR，根据信息去检索Item
  - A_信息：Query、OS、Browser、PID；
  - B_广告：Bidword、Title；
  - A&B间的匹配；



# L2M

- 建模：

$$eCPM(A,B) = CTR(A,B) * Bid(B)$$

- CTR：
  - 难点：A&B
  - 组合特征，泛化新能差；
  - 样本哪里来；
- Bid：

$$Bid(B) = \sum_{A_i \in A} Bid(B, A_i)$$

方法：
1，稀疏特征：交叉特征；
2，稀疏特征：隐因子模型；
3，稠密特征，非线性模型；

# 隐因子模型

- Latent Model
  - LFM：

$$\hat{r}_{u,m} = p_u^T q_m = \sum_{k=1}^{K} p_{uk} \cdot q_{mk}$$

  - Bias-LFM：

$$\hat{r}_{u,m} = b_u + b_m + p_u^T q_m = b_u + b_m + \sum_{k=1}^{K} p_{u,k} \cdot q_{m,k}$$

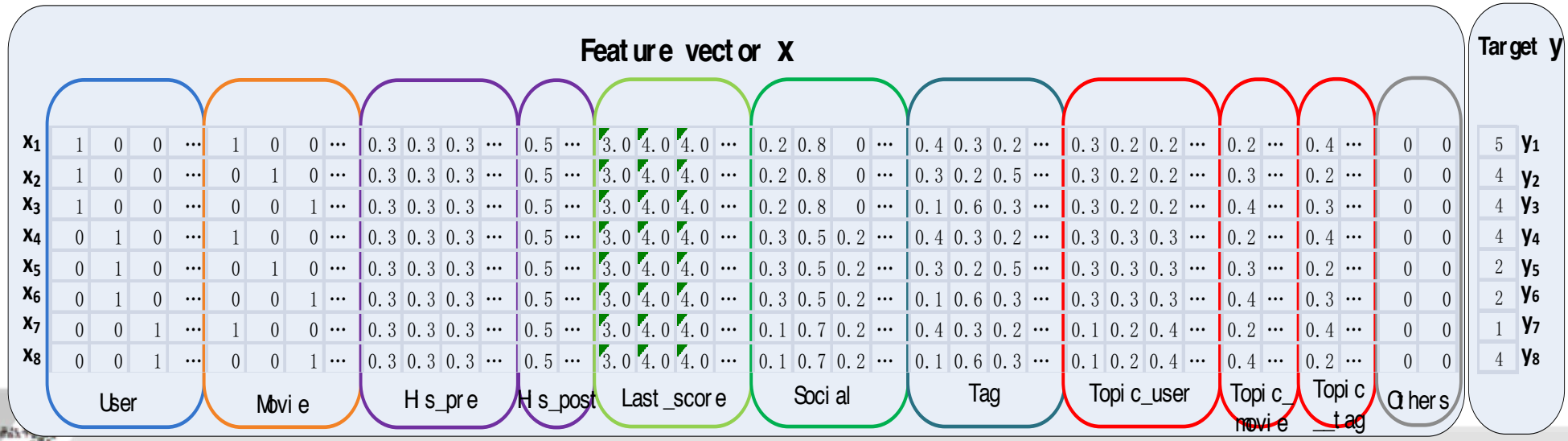
  - Time Bias-LFM：

$$\hat{r}_{u,m} = u + b_u + b_m + b_t + p_u^T q_m + x_u^T y_t + s_i^T z_t + \sum_f g_{u,f} h_{if} l_{tf}$$

# 隐因子模型

- Fatorzation Machine：

$$\hat{y}(x) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} < v_i, v_j > x_i \cdot x_j$$



# L2M：模型

- 预测：

$$\hat{y}(x) := \frac{1}{1+\mathrm{e}^{-sum}}$$

$$sum = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} < v_i, v_j > x_i \cdot x_j$$

- 损失函数：$L(x) = \|w\| + C * Log(1 + e^{-y*sum})$

- 优化：
  - Adapative Learning Rate；
  - Hyper Parameter Learning；

# L2M： 整体流程

- 探索（EXPLORATION）：
  - 线上随机一部分流量，收集样本；
- 利用（EXPLORTATION）：
  - 训练模型；
  - 挖掘触发供线上使用；

THANKS

SeaqueMedia 盛拓传媒 ITPUB ChinaUnix ITPUB