中国联通
应用商店运营中心

基于用户画像的大数据挖掘实践

杨步涛

2014年11月

纲要

1 沃商店定位

2 沃商店大数据体系架构

3 用户画像建设

4 个性化推荐

5 广告

6 用户画像的其他应用实例

沃商店定位

渠道聚合

- 2013年中国手机应用分发总量快速上升，其中应用商店的分发量占比超过80%；
- TOP10渠道占总分发量的90%。

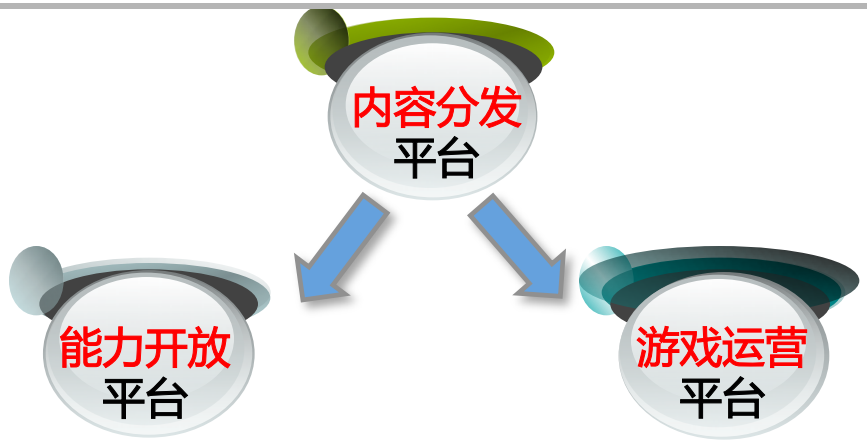
手游爆发

- 2013年中国手机游戏市场近100亿元，2014年预计将达到180亿元。
- 多款优质手游月流水超过5000万元。

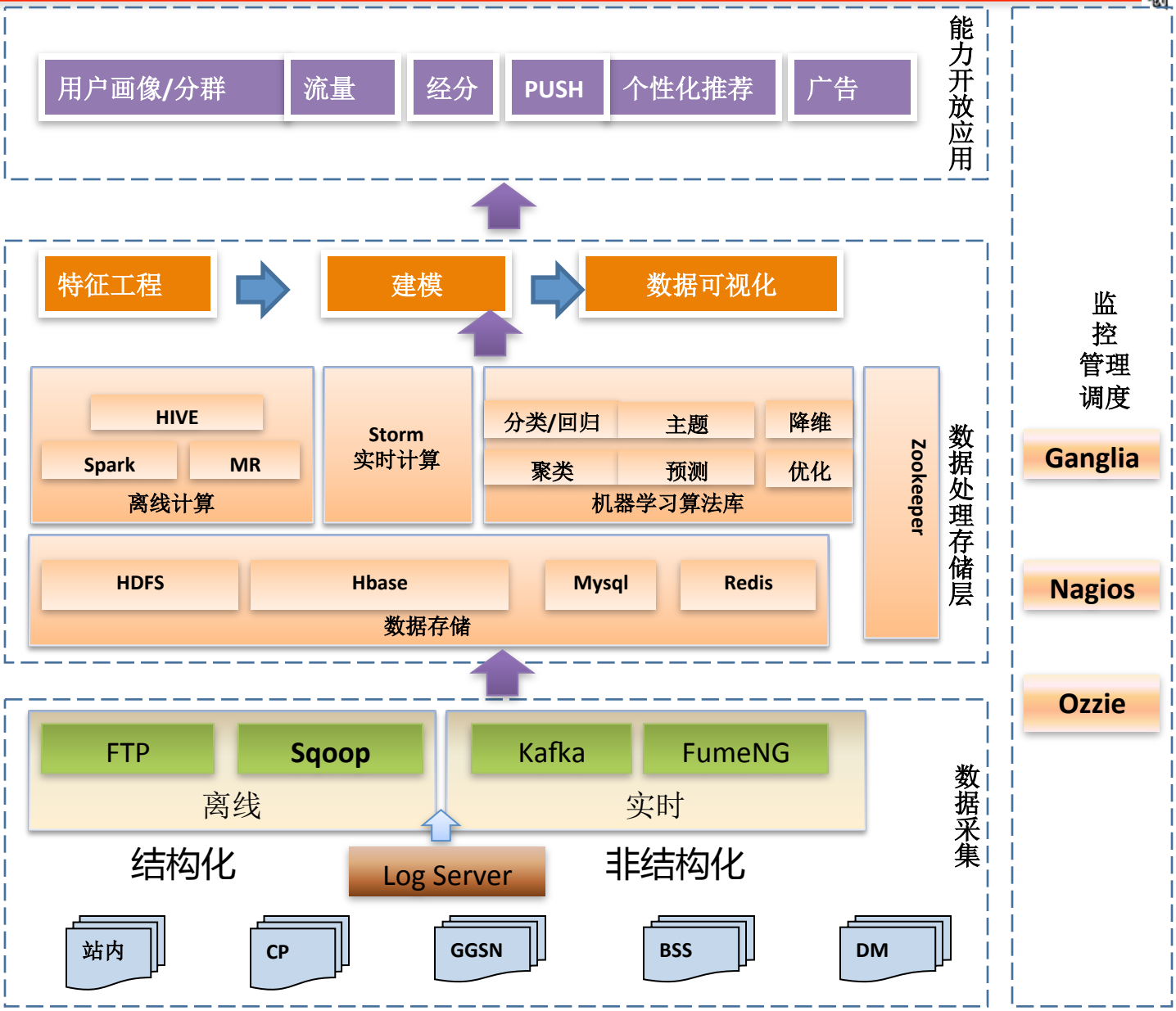
支付变革

- 运营商通信账户SDK支持APP应用内付费，提升付费转换率，从不足5%提升至20%以上。
- 话费支付的便捷优势机遇期短暂仅1-2年，移动互联网支付的替代转瞬即至。

沃商店定位



沃商店大数据架构体系


能力开放应用
用户画像/分群
流量
经分
PUSH
个性化推荐
广告
特征工程
建模
数据可视化
HIVE
Spark
MR
离线计算
Storm
实时计算
分类/回归
主题
降维
聚类
预测
优化
机器学习算法库
Zookeeper
数据处理存储层
HDFS
Hbase
Mysql
Redis
数据存储
FTP
Sqoop
离线
Kafka
FumeNG
实时
数据采集
结构化
Log Server
非结构化
站内
CP
GGSN
BSS
DM
监控管理调度
Ganglia
Nagios
Ozzie

特征工程—用户画像

经分预测 | 广告 | 流量 | 智能推荐 | PUSH

数据开放

沃商店、CP

结构化

语义挖掘

知识库 自我学习 行为标注引擎

网络爬虫引擎

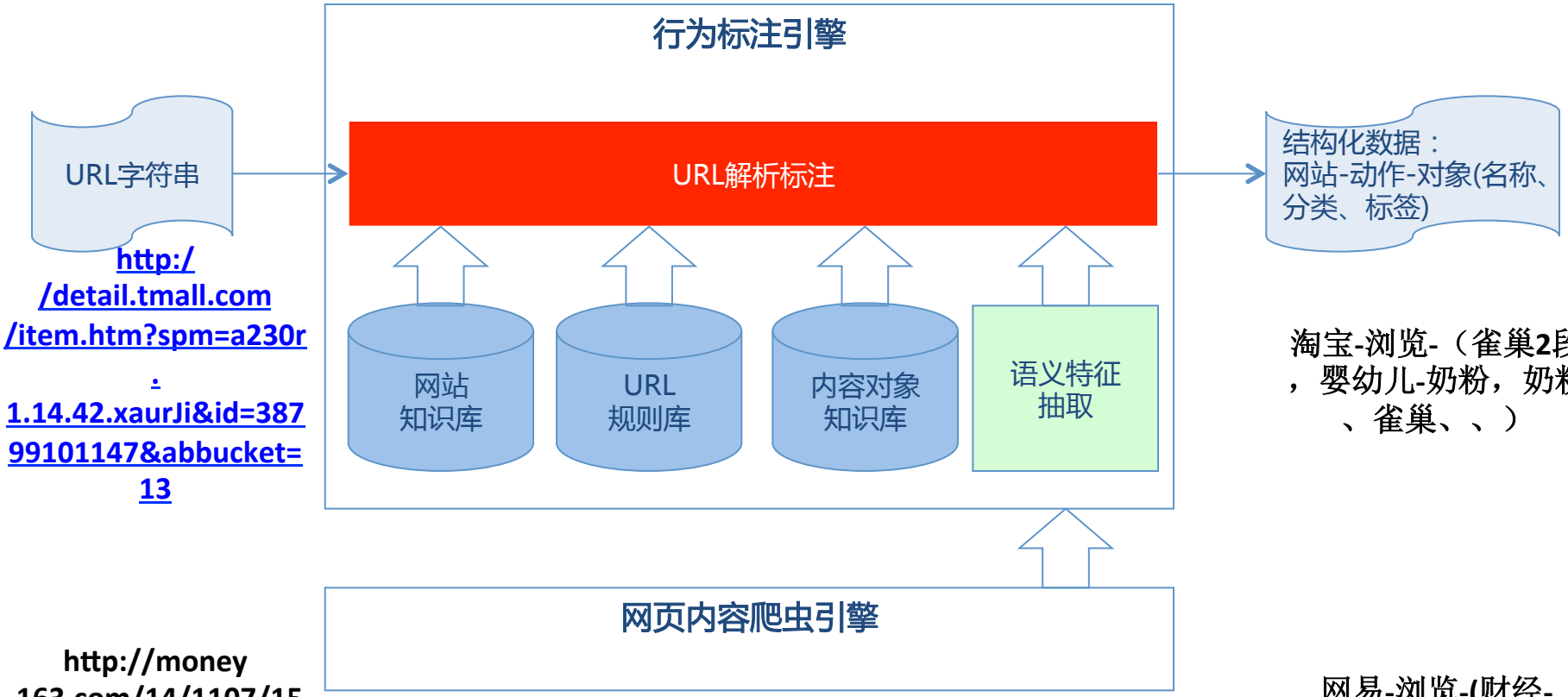
非结构化

手机搜狐 手机凤凰网 sina新浪网 VANCL 淘宝网

基本维度	性别、年龄、地域、终端、网络、渠道、机型...
兴趣、标签维度	影音、生活、社交、金融、阅读、儿童亲子...
商业维度	转化、新老用户、频次、付费次数、ARPU、通信消费、信用等级

用户画像

用户画像—行为标注



http://detail.tmall.com/item.htm?spm=a230r.1.14.42.xaurJi&id=38799101147&abbucket=13

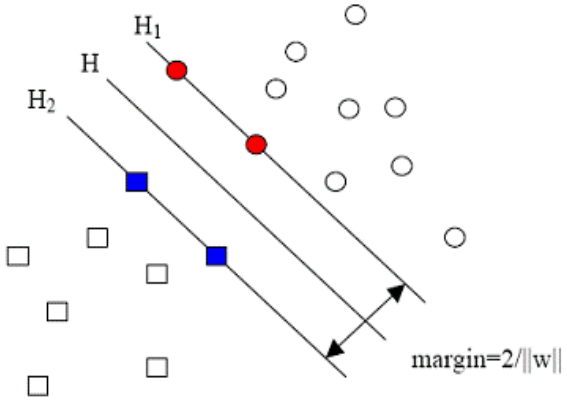
淘宝-浏览-（雀巢2段，婴幼儿-奶粉，奶粉、雀巢、、）

http://money.163.com/14/1107/15/AAF5A98J00254IU4.html

网易-浏览-(财经-证券，、、)

用户画像—网页文本分类

- 特征预处理、特征筛选(降维)
- 支持向量机SVM
  - 结构风险最优化
  - 非线性(核函数、松弛变量)
  - 1对1方式多分类支持
- 评估：准确率、召回率、F1

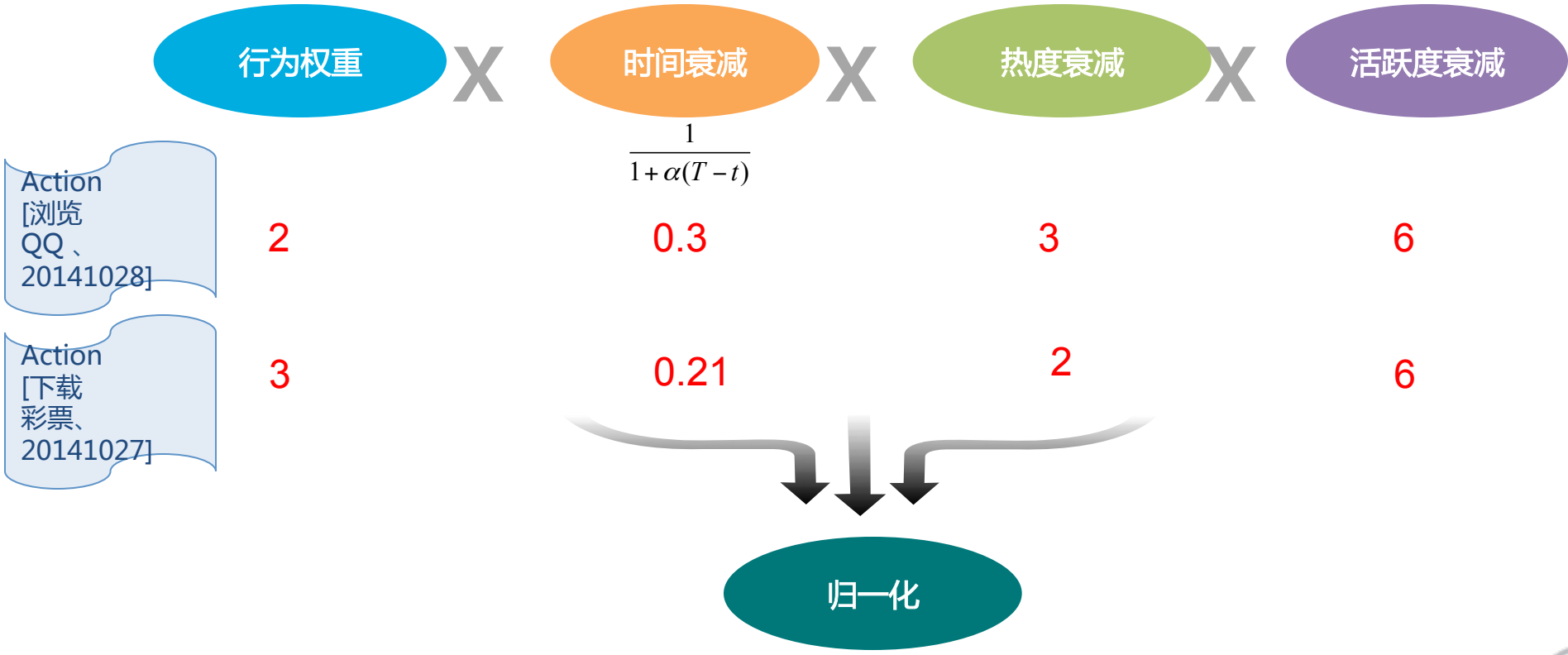


$$\min \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \zeta_i$$

$$\text{subject to} \quad y_i[(wx_i)+b] \geq 1-\zeta_i \quad (i=1,2,\ldots,l)$$

算法	准确率(P)	召回率(R)	F1
朴素Bayes	85%	86.2%	85.5
SVM	92%	93%	92.4

用户画像—兴趣建模


行为权重
X
时间衰减
X
热度衰减
X
活跃度衰减
$\frac{1}{1+\alpha(T-t)}$
Action [浏览QQ、20141028]
2
0.3
3
6
Action [下载彩票、20141027]
3
0.21
2
6
归一化

用户画像应用实践

- 个性化推荐
- 广告
- 信用等级分群
- 用户流失预警
- 游戏潜在用户群体筛选
- 异常监控分析

推荐—应用场景

首页推荐

应用详情推荐

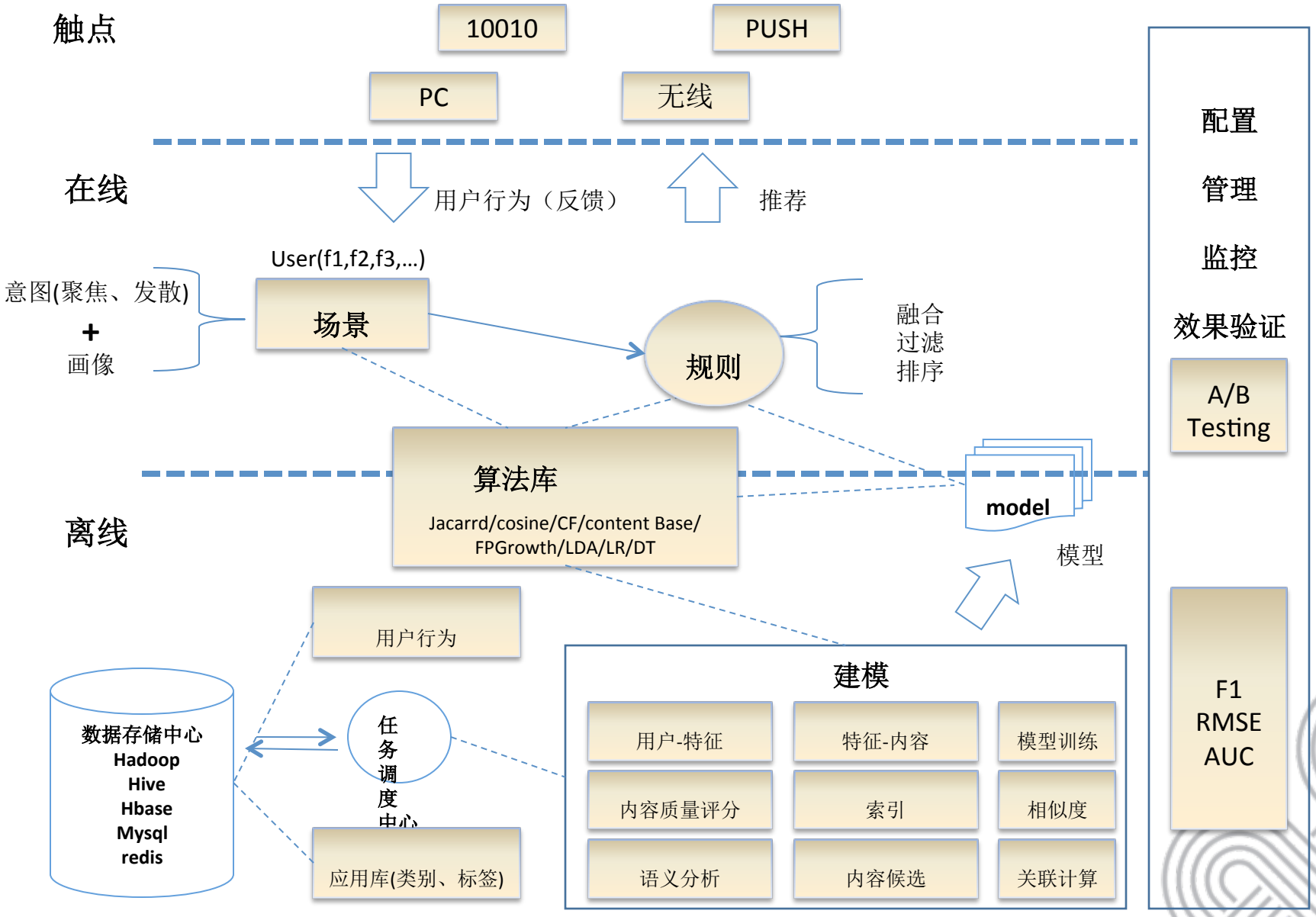
用户粘性

转化率

广告

猜你喜欢

个性化推荐—平台架构


触点
10010
PUSH
PC
无线
在线
用户行为（反馈）
推荐
User(f1,f2,f3,...)
意图(聚焦、发散)
+
画像
场景
规则
融合
过滤
排序
算法库
Jacarrd/cosine/CF/content Base/
FPGrowth/LDA/LR/DT
model
模型
离线
用户行为
数据存储中心
Hadoop
Hive
Hbase
Mysql
redis
任务调度中心
应用库(类别、标签)
建模
用户-特征
特征-内容
模型训练
内容质量评分
索引
相似度
语义分析
内容候选
关联计算
配置
管理
监控
效果验证
A/B Testing
F1
RMSE
AUC

个性化推荐—算法

Content-Based

Category

冷启动

推荐精度

ItemBased-CF

$$p(u,j)=\sum_{i\in N(u)\cap S(j,K)} w_{ji}r_{ui}$$

新颖

$$w_{ij}=\frac{|N(i)\cap N(j)|}{\sqrt{|N(i)||N(j)|}}$$

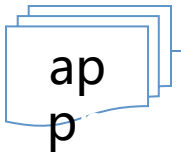
稀疏

$$w_{ij}=\frac{\sum_{u\in N(i)\cap N(j)}\frac{1}{\log 1+|N(u)|}}{\sqrt{|N(i)||N(j)|}}$$

活跃用户

Model-Based

聚类模型



语义分析

LDA

Topic分布

基于KL距离
推荐语义相关应用

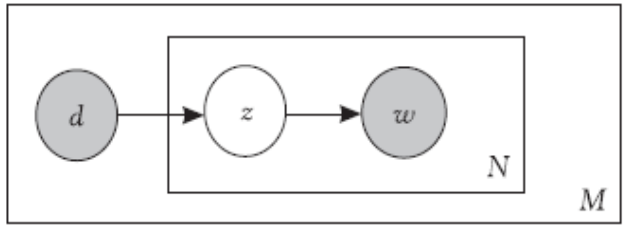
$$D_{\mathrm{KL}}(P\|Q)=\sum_i \ln\left(\frac{P(i)}{Q(i)}\right)P(i).$$

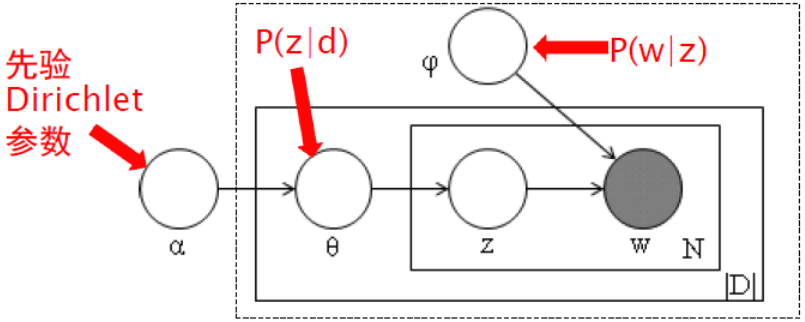
来源融合

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

基于语义相关推荐 —应用主题抽取

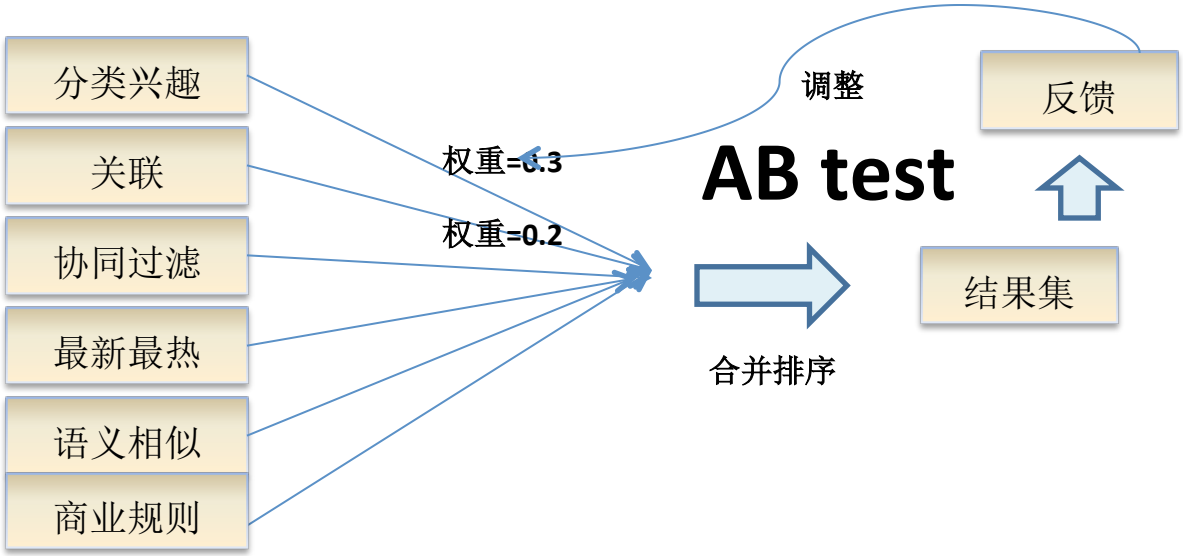
- PLSA
  - EM
  - 无p(z|d)的生成概率模型
  - 容易过拟合
- LDA
  - Gibbs
  - 参数少，过拟合风险小
  - 新文档处理能力强





算法融合—调整前

如何确定各个模型、特征的权重？



人工对权重的调整，很难把控
新加入特征难以快速设置特征

算法融合—改进后

Logic Regression

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

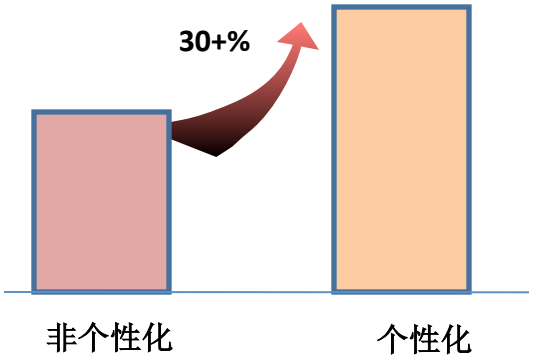
权重系数： $h_\theta(x)=\theta_0x_0+\theta_1x_1+\theta_2x_2+\ldots+\theta_nx_n$

候选集	model1	model2	model3			score
App1	0.2	0.54				0.7
App2		0.32	0.6			0.5

学习隐藏在用户群体行为背后的规律

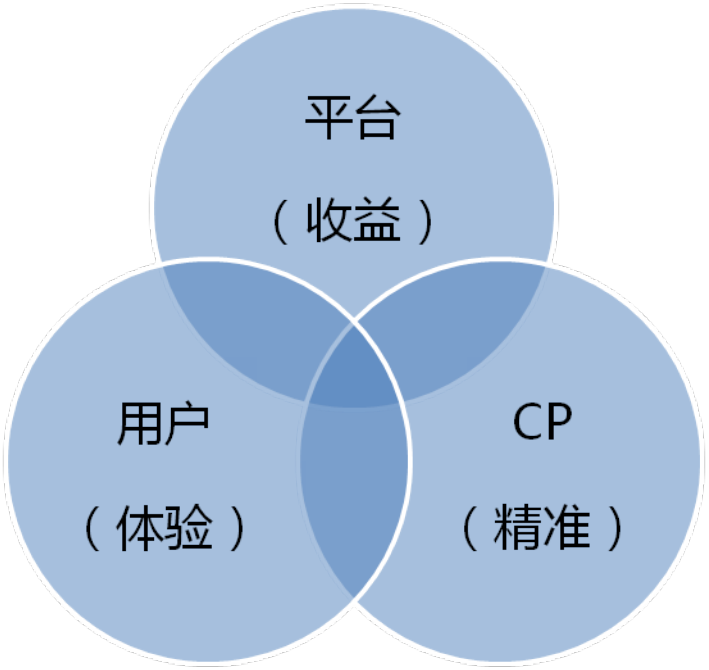
个性化推荐—效果评价

PV转化率(CTR*CVR)，效果提高30%



个性化推荐的下载量占比21%

广告



- 公式：ctr*Bid

  ctr=click/PV

- 缺点：冷启动、缺少个性化诉求

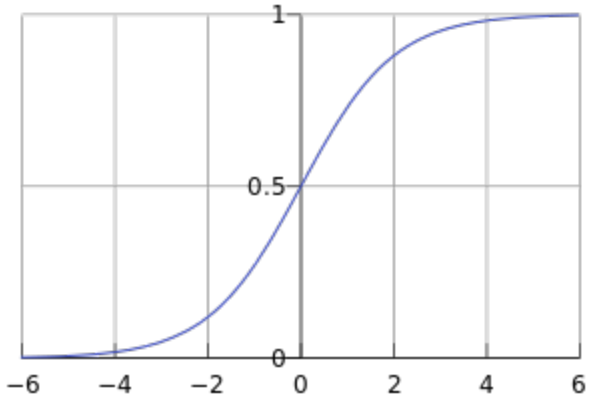
广告—改进算法

- 公式：pCTR*Bid
- pCTR：p(click|ad,user)
- 基于LR的点击预估模型

  点击=1，不点击=0

  点击的概率

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



广告—特征抽取

- 特征

  用户profile(活跃度、性别、年龄、标签)、
  广告(广告质量、历史点击率、新颖性)、
  CP、
  用户和广告交叉主题特征

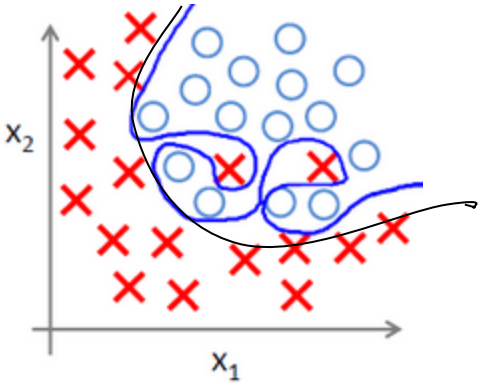
广告—样本和特征处理

- 样本选择
  - 去噪、样本抽样
- 特征处理
  - 归一化
  - 离散化、交叉
  - 泛化能力

    正则化(惩罚)

    L1，使得大量无效特征权重为0

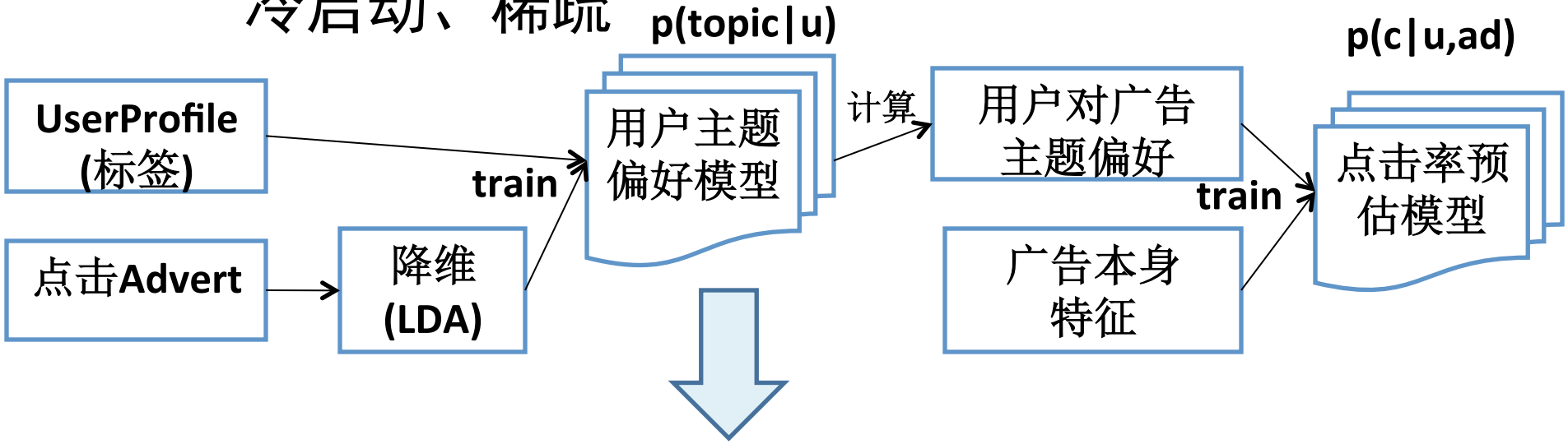
    L2



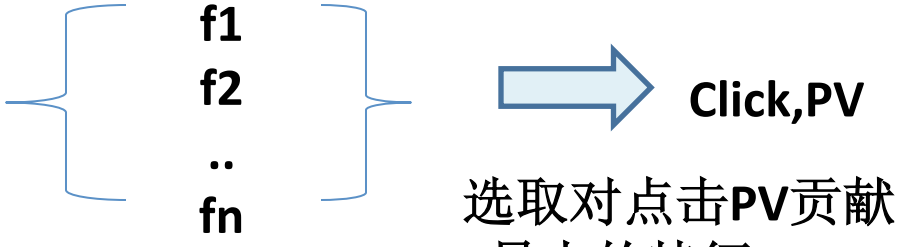
广告—特征处理(续)

■ 特征处理

冷启动、稀疏



用户标签特征筛选(降维)



选取对点击PV贡献
最大的特征TopN

广告—LR参数优化

- 参数估计优化：

  L-BFGS

  line search确定步长，无须手动选择，

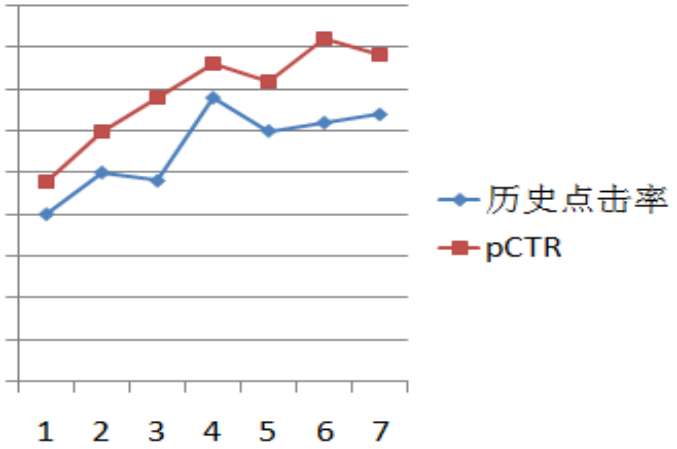
  利用有限内存近似BFGS，

  利用历史值和梯度寻找当前方向(Two loop)，

  实现快速迭代

广告—效果评估

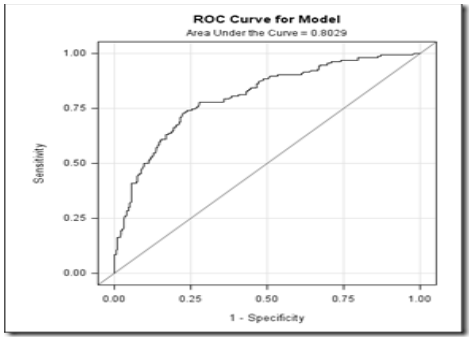
■ 产品层面



■ 算法层面

AUC

对于CTR高的广告，
预测的是否也高？



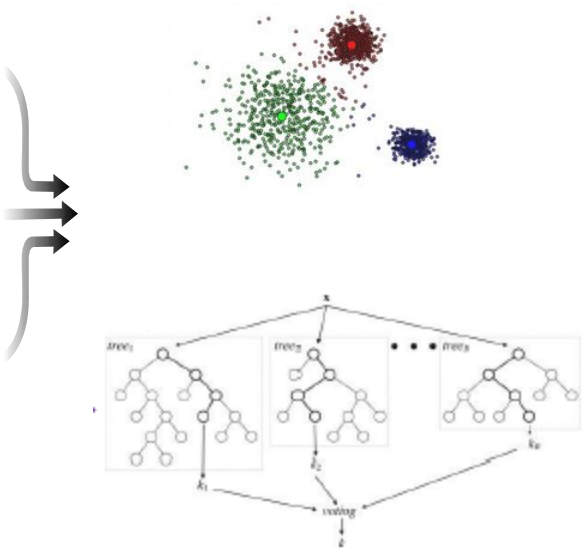
算法、数据、人机交互

用户信用等级分群

出账信息	入网时间
入网套餐价格	预付/后付
月流量消耗	号码归属地

支付	消费频次
消费金额	支付成功率
	……

欠费情况	*客服情况*



L1

L2

L3

…

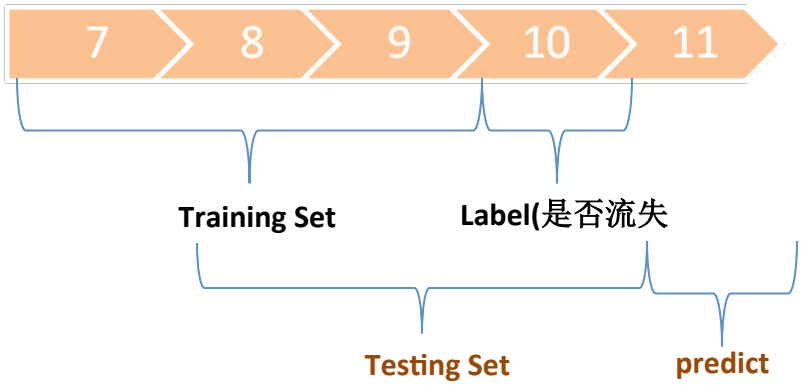
LN

基本算法

- **K-means**
- **RF**

用户流失预警

■ 模型



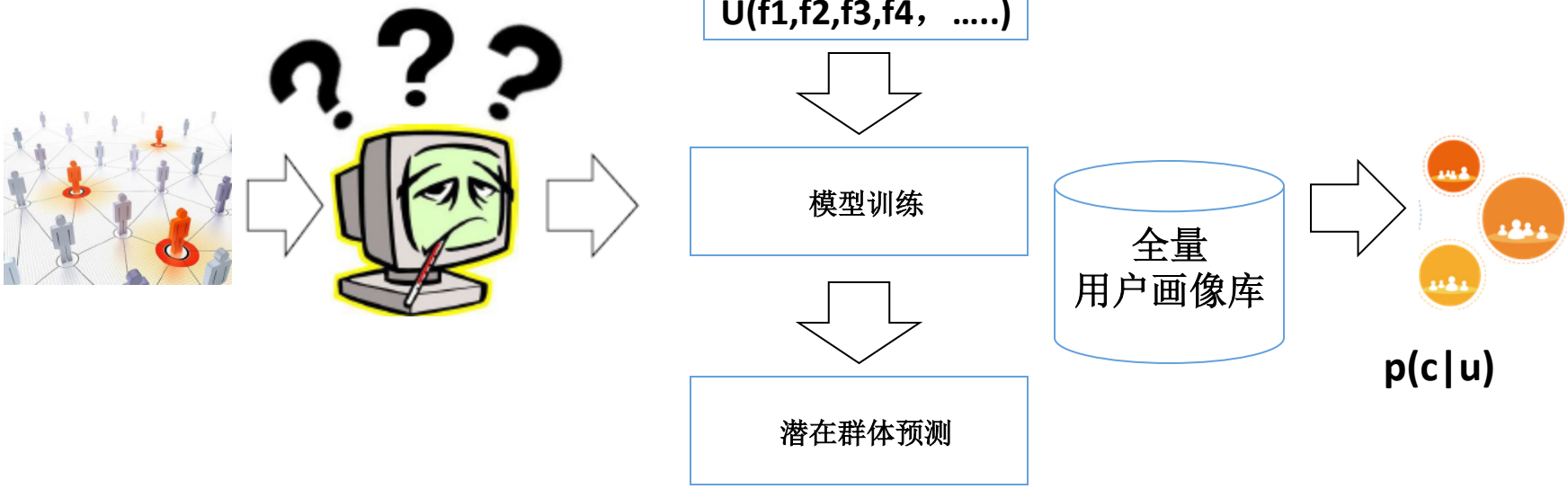
10月份流失的用户，
分析其前3个月的行为数据

■ 特征

活跃度、登陆情况、下载情况、预装机情况、机型、....
➔潜在的流失用户
针对可能流失的用户做PUSH推广活动

基本算法
●GBDT

游戏潜在用户群体分析

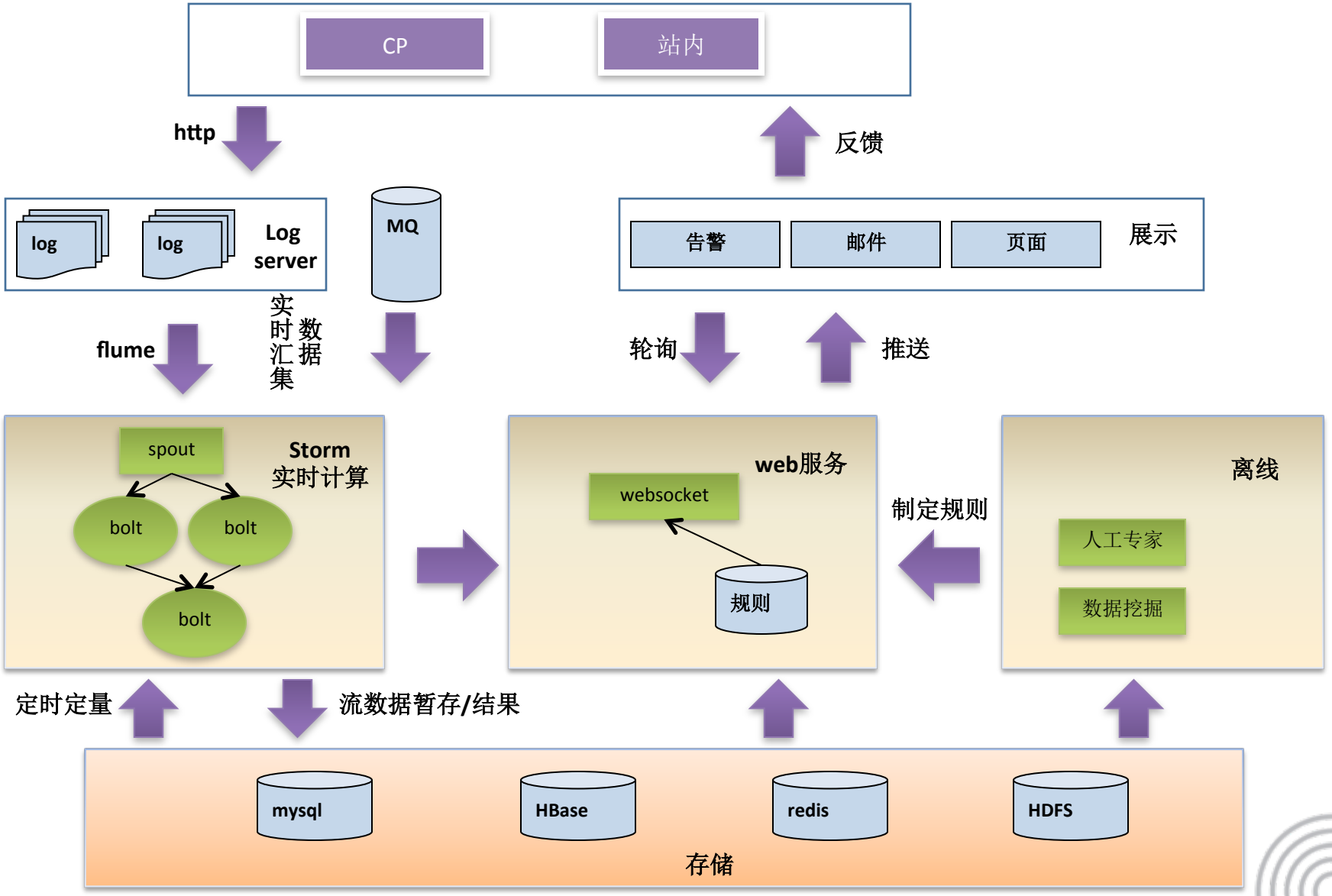


以在应用或者游戏中产生转化(注册、付费)的这些用户作为训练正样本，结合用户特征进行模型训练，从用户画像库中筛选出潜在的用户群体，推荐给CP，通过PUSH做相关的营销活动

基本算法

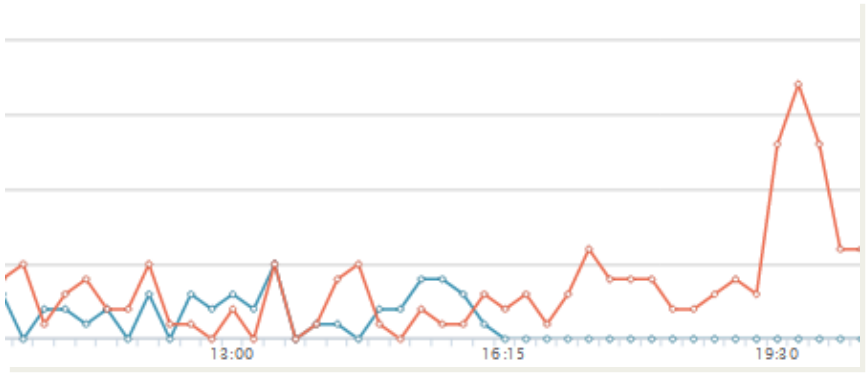
- Logic Regression

异常监控—Storm流计算

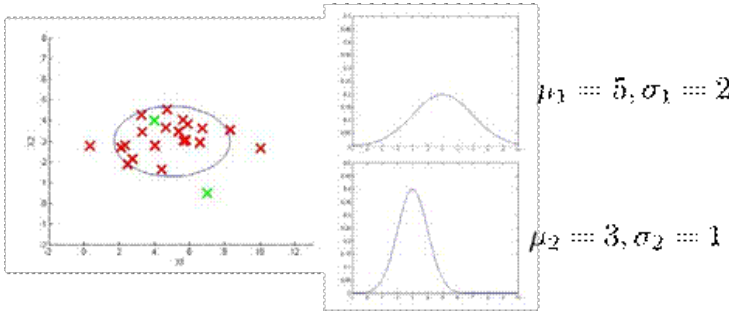


异常监控和分析—离群点分析

■ 异常检测



$$p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}\right)$$



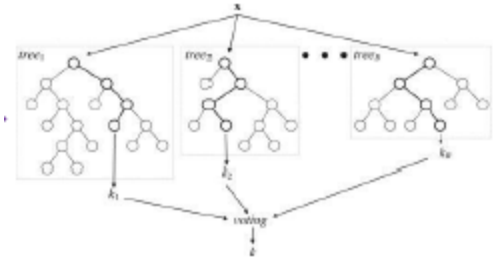
■ 刷机诊断

基于规则(IF ELSE)：依赖经验，调整繁琐，准确度低

基于模型：

利用用户刷机的一些行为特征和数据

进行模型训练，结合模型来判断当前是否刷机



招聘

我的blog：

http://blog.csdn.net/yangbutao

我们招聘：

Hadoop/Hbase/Spark开发

算法工程师

数据挖掘工程师

...

China unicom中国联通
创新·改变世界

敬请期待：
2015中华数据库大会
时间：2015.05.16
报名时间：2015.02.14
报名网址：meeting.zhdba.com

联系我们：
联系人：朱小姐
联系电话：136 5197 9898
联系QQ：378091820