Geekbang 极客邦科技 InfoQ

QCon 全球软件开发大会【北京站】2016

广告平台中用户画像和标注噪声处理的实践

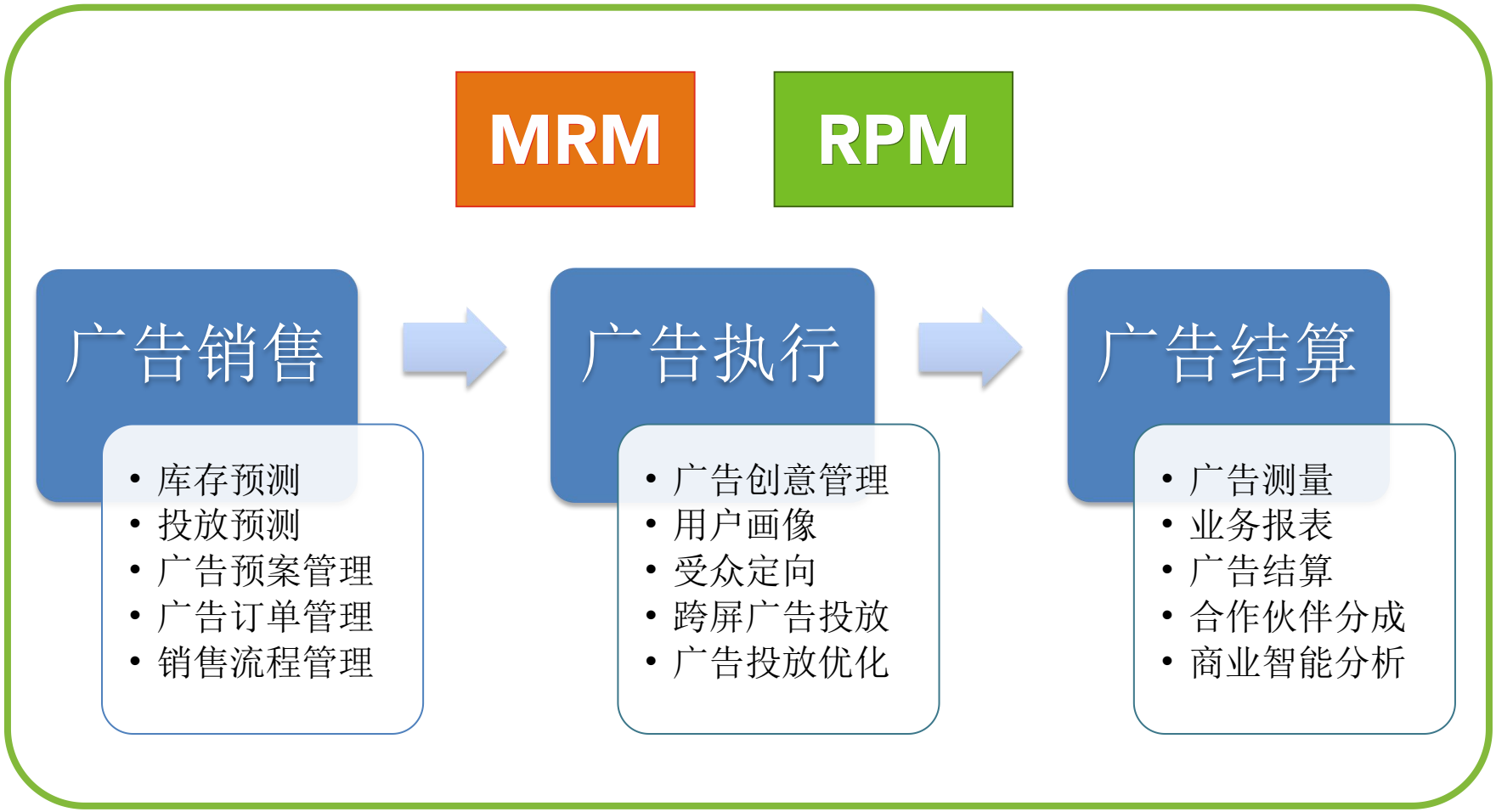
飞维美地信息技术(北京)有限公司
童有军 yjtong@freewheel.tv

Intemational Software Development Conference

主要内容

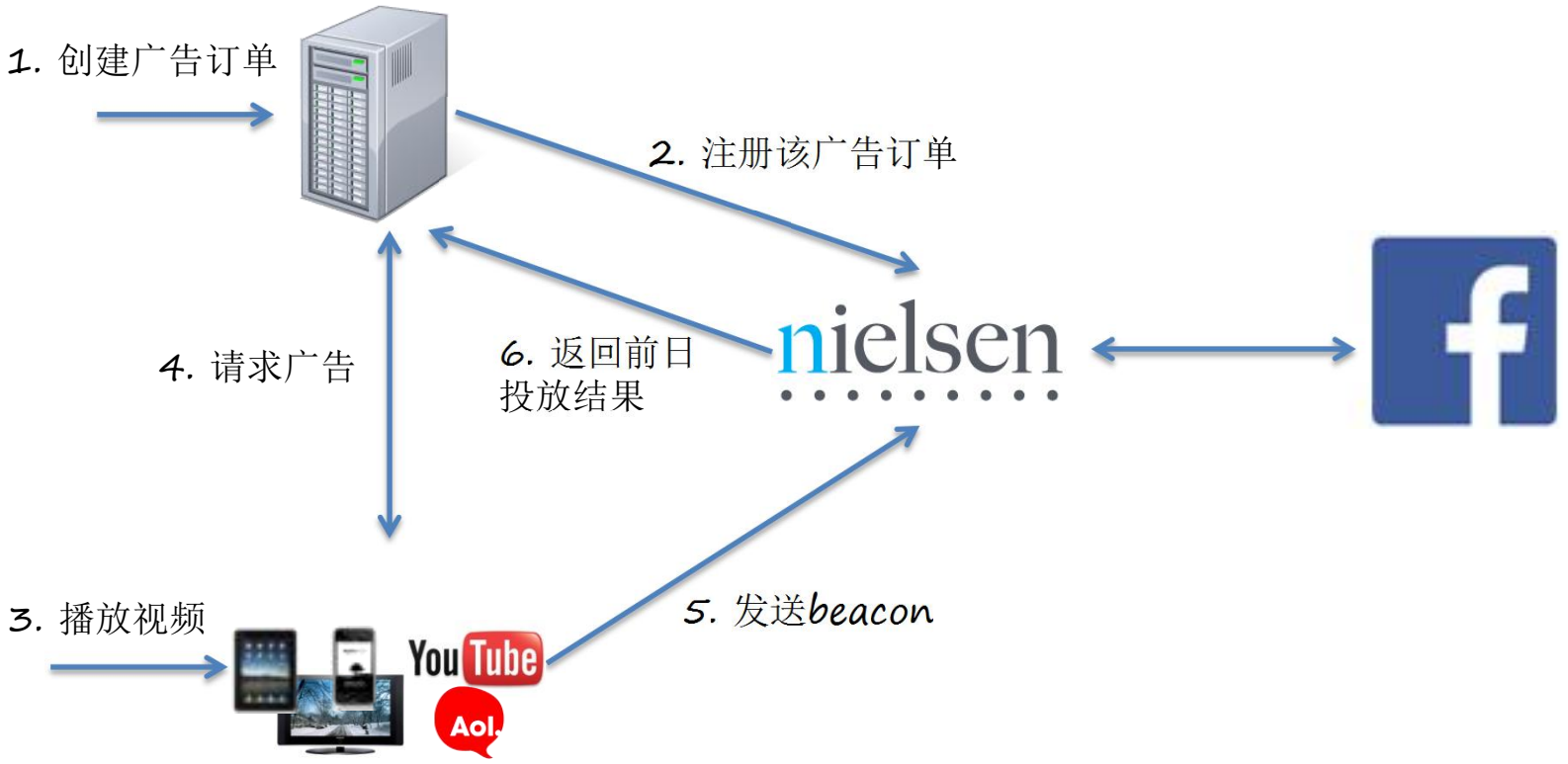
- 背景介绍
- 用户画像
- 标注噪声
- 系统架构

我们的业务



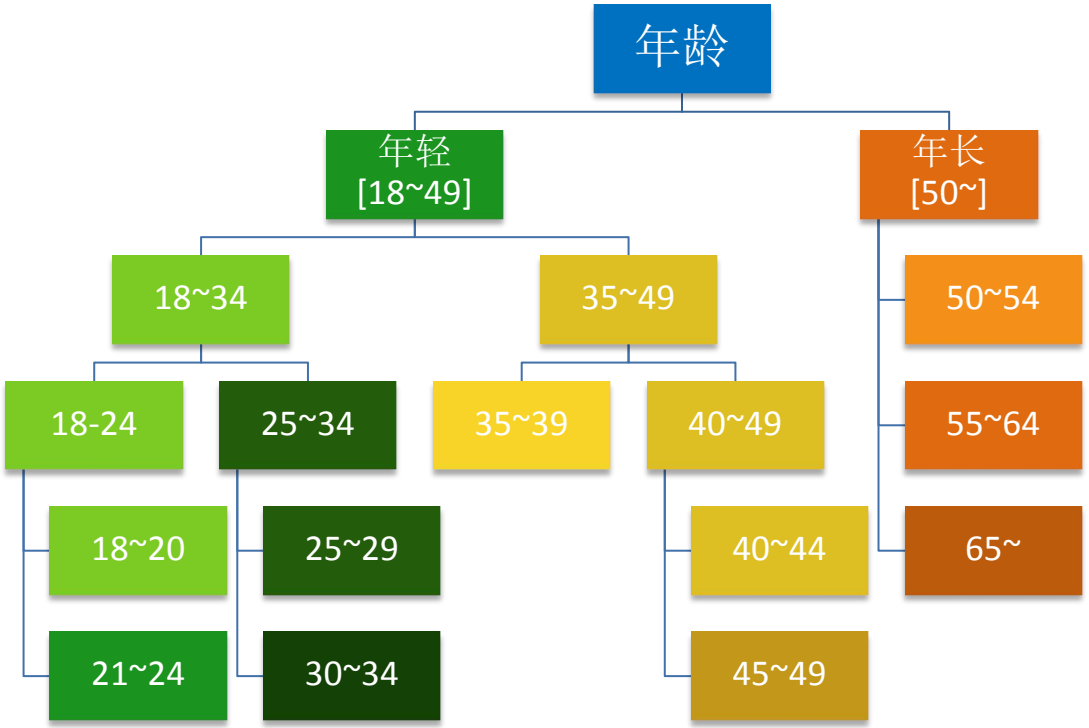
单日投放近10亿次广告，生成2TB广告投放数据

数字收视率测量流程


1. 创建广告订单
2. 注册该广告订单
4. 请求广告
6. 返回前日
投放结果
nielsen
3. 播放视频
5. 发送beacon
YouTube
Aol.

用户画像


性别


年龄
年轻
[18~49]
年长
[50~]
18~34
35~49
50~54
18-24
25~34
35~39
40~49
55~64
18~20
25~29
40~44
65~
21~24
30~34
45~49

挑战

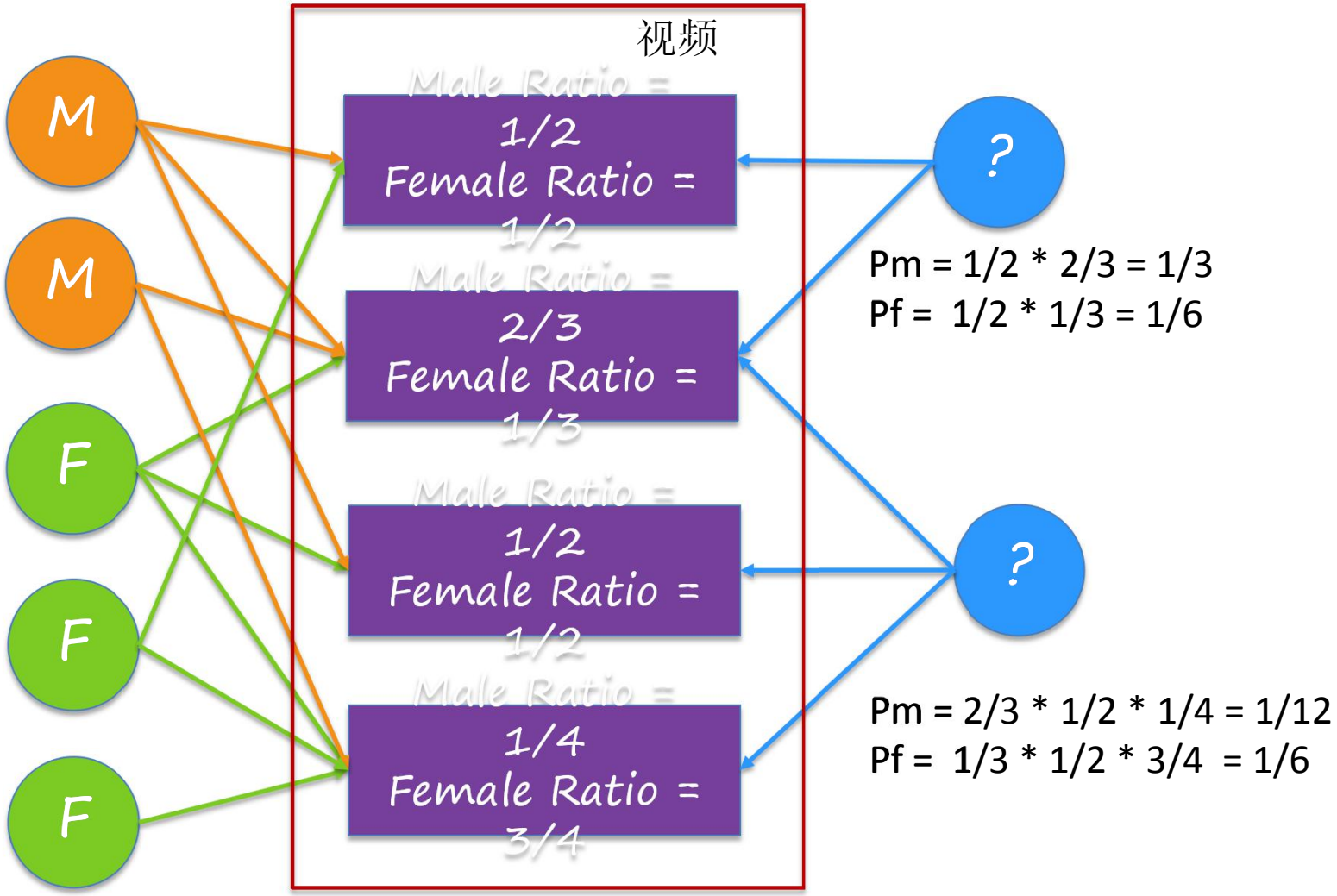
缺少较好的标注来源

- 第三方按批次返回结果
- 其他的标注集质量低。例如：在性别上，BlueKai的标注数据的准确率不足60%

如何用特征描述用户

- 对视频信息掌握的有限

生产标注


视频
M
M
F
F
F
Male Ratio = 1/2 Female Ratio = 1/2
Male Ratio = 2/3 Female Ratio = 1/3
Male Ratio = 1/2 Female Ratio = 1/2
Male Ratio = 1/4 Female Ratio = 3/4
?
?
Pm = 1/2 * 2/3 = 1/3
Pf = 1/2 * 1/3 = 1/6
Pm = 2/3 * 1/2 * 1/4 = 1/12
Pf = 1/3 * 1/2 * 3/4 = 1/6

贝叶斯方法

- 基本公式

$$Pr(c|u_i) \propto Pr(c|\{v\})$$
$$\propto Pr(\{v\}|c)Pr(c) \propto \prod_j^k Pr(v_j|c)\, Pr(c)$$
$$= \frac{\prod_j^k Pr(c|v_j) Pr(v_j)}{Pr(c)} Pr(c)$$
$$\propto \prod_j^k Pr(c|v_j)$$

u: 用户; v: 视频; c: 类别

[{jianh,hjzeng,etc}@microsoft.com]

标注结果

- 分类

$$If\ \frac{\Pr(c1|u)}{\Pr(c2|u)} > 1, then\ u\ is\ c1;\ else\ u\ is\ c2$$

- 选择置信度高的部分(60%)，作为初始标注集合

类别	男性	女性	年轻	年长
准确率	82%	80%	75%	82%

特征工程

Mountain View
CA, US

Ubuntu 14.04

Chrome

2016-03-02
07:26 am: 1次观看
07:30 am: 1次观看
12:40 pm:1次观看
20:00 pm:1次观看
21:18 pm:1次观看
23:11 pm:1次观看
00:21 pm:1次观看

静态特征

行为记录

观看历史

规则特征

- 离散特征

特征	例子	来源	转化
地域	CA ， US	IP	One-Hot
设备，操作系统，播放器(浏览器)	Apple，iOS，Chrome	User Agent	One-Hot

- 连续特征

视频	划分
Short	时长 <= 5分钟
Middle	5分钟 <时长<=20分钟
Long	时长 > 20分钟

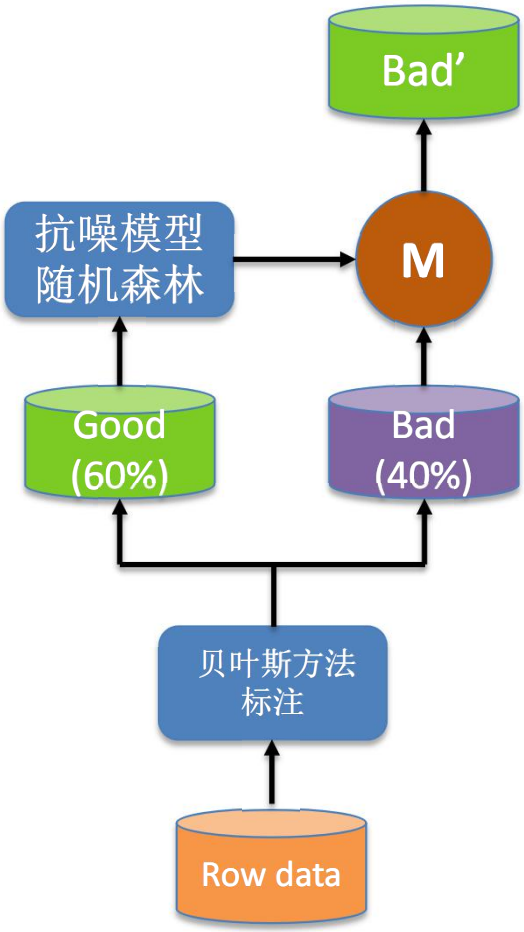
观看时间	划分
Early Morning	6 AM – 10 AM
Daytime	10 AM – 4 PM
Early Fringe	4 PM – 7 PM
Primetime	7 PM – 1 AM
Graveyard	1 AM – 6 AM

视频特征

- 基于视频描述文本的标签分类
  - 视频提供方会对一部分的视频提供标签信息
    - 例如：*Animation, Business, Comedy, Entertainment, News, Sports…*
  - 朴素贝叶斯构建特征
    - 分词后，根据term在正负样本中的分布来构造特征
    - Bigram
  - 基于Spark的逻辑回归
    - 1 vs 其他
- LDA
- 命名实体

[{sidaw,manning}@stanford.edu]

用户画像



准确性	贝叶斯方法 (覆盖60%)	随机森林 (覆盖100%)
男性	82%	72%
女性	80%	69%
年轻	75%	66%
年长	82%	71%
问题	标注集的准确性依然不够，标注集合中有大量噪声	

标注噪声

- 需要进一步消除噪音
  - 分类树的优势：一层层的优化
  - 主要的消噪方法
    - Boosting
    - Bagging
      - Cross-Validation
      - 有放回的抽样
    - 半监督的方法
  - 模型性能的评测
    - 线下标注的纠正比率
    - 线上实际准确率

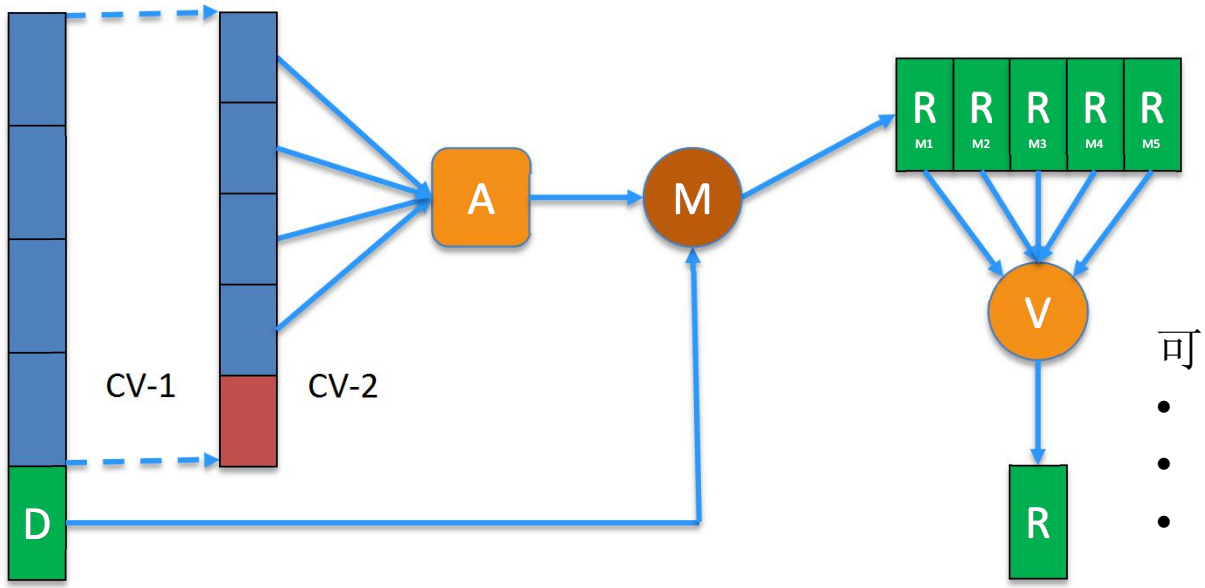
Boosting

- AdaBoost
  - 思路
    - 在Boost算法的训练过程中，噪声数据有得到更大的权重的趋势
    - 在多轮迭代后，获得高权重的数据更有可能为噪声
  - 结果

分类	纠正比率（剔除比率）	准确率
年轻	23%	80%
年长		81%

[{yoav, schapire}@research.att.com]

Bagging



可调参数：

- CV-2的方法
- A:算法包的组成
- V: 投票的策略

[verbaeten,vanassche]

不同的Bagging

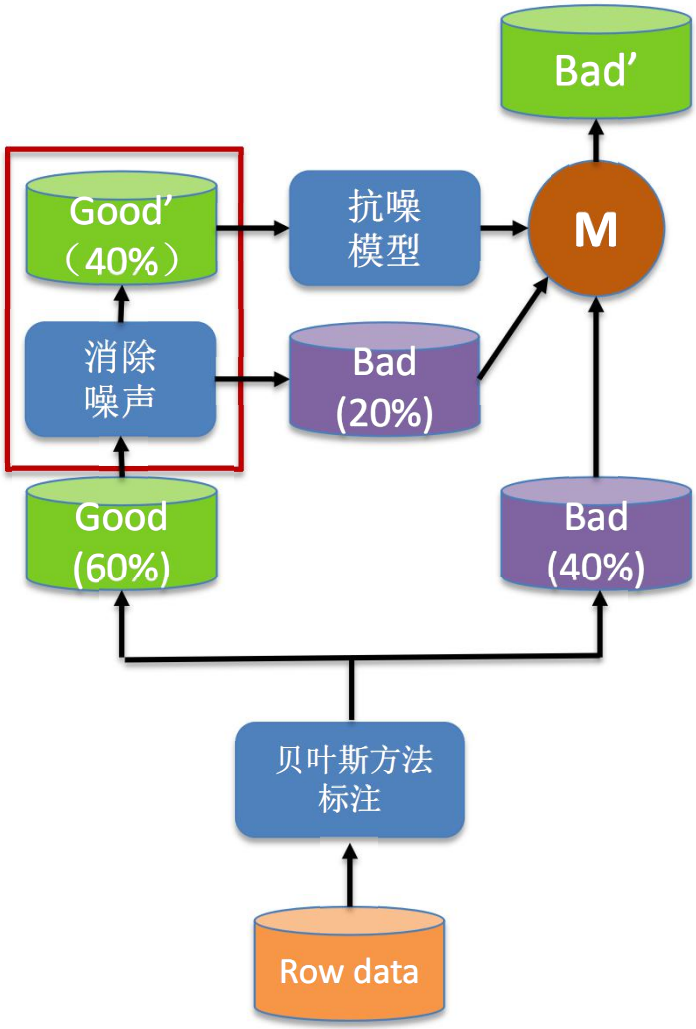
- CV-2的方法
  - 5折cross-validation，每个子集间相互独立
  - 放回抽样，每次随机抽取80%的数据进行训练
- 算法包的组成
  - 5-模型算法包：1个决策树，3个kNN(k＝[1,3,5])，1个逻辑回归
  - 9-模型算法包：9个决策树
- 投票的策略
  - 一致性投票
  - 大多数投票

Bagging的比较

- 评测结果
  - 年龄维度

模型编号	CV2方法	算法包	投票方法	纠正比率	准确率	
					年轻	年长
CV_5_C	5-fold	5-model	一致性	5.54%	NA	NA
Ba_5_C	Bagging	5-model	一致性	5.57%	NA	NA
Ba_5_M	Bagging	5-model	大多数	15.48%	80.78%	84.93%
CV_5_M	5-fold	5-model	大多数	16.03%	82%	85%
CV_9_M	5-fold	9-model	大多数	32.36%	85.96%	86.43%

消除噪声



	初始标注集		消除噪声后	
准确率	年轻	年长	年轻	年长
	75%	82%	85%	85%
覆盖面	60%		40%	
问题	Boosting和Bagging的消噪方法都会损失数据，需要扩大召回			

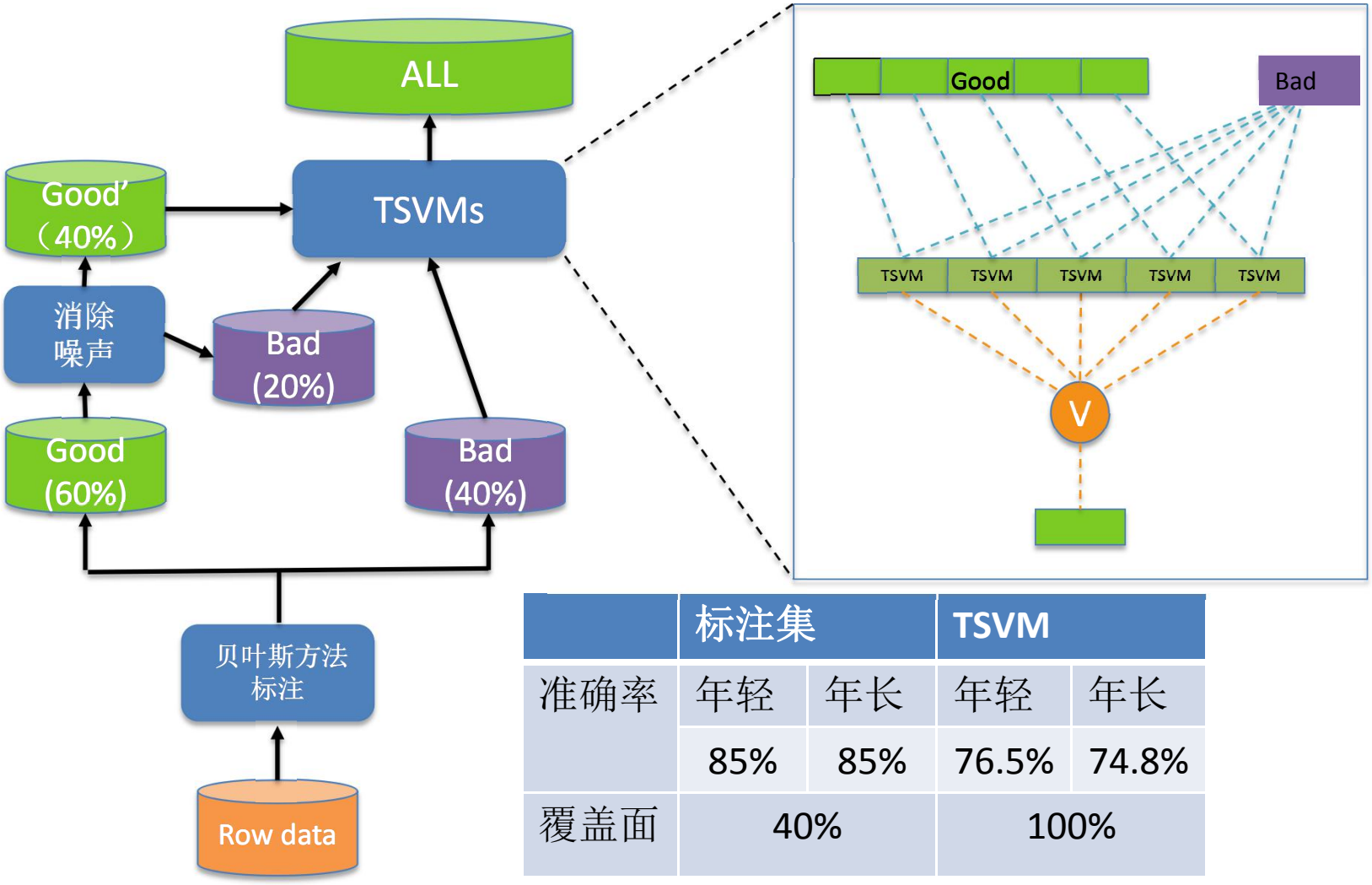
扩大召回

- 利用随机森林扩大召回

类别	年轻	年长
准确率	71%	72%
覆盖面	100%	

- 解决方案
  - 半监督学习
  - TSVM(Transductive Support Vector Machines)

TSVM

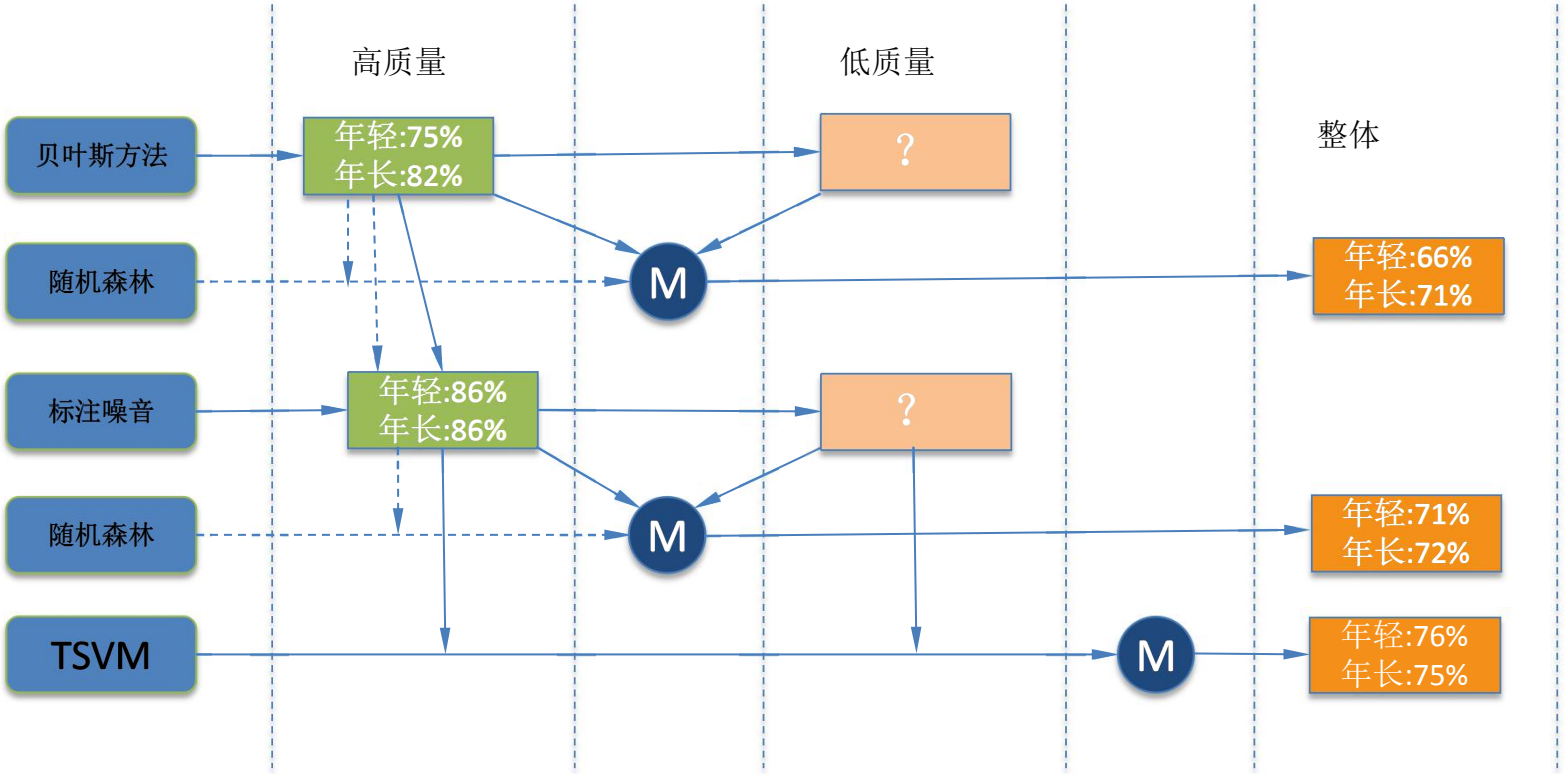


	标注集		TSVM	
准确率	年轻	年长	年轻	年长
	85%	85%	76.5%	74.8%
覆盖面	40%		100%	

[Sindhwani, Keerthi]

标注噪声总结

- 方法的演进

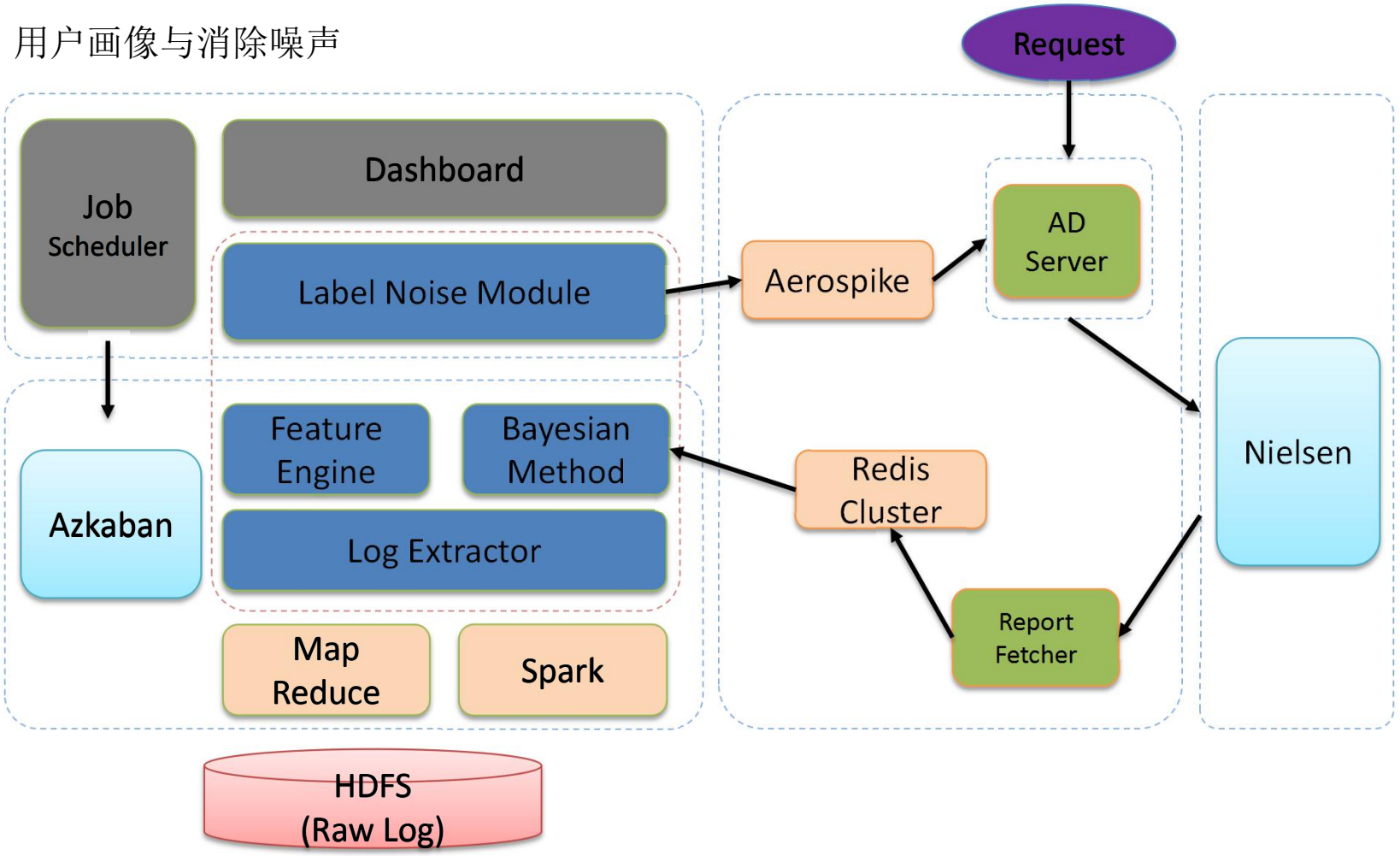


一些经验

- 需要注意的几点问题
  - 特征的问题
    - 模糊特征
    - 特征筛选的经验
  - 模型的选择
    - 抗噪模型
  - 离线参数的调试
    - 训练误差的提示

系统架构

- 用户画像与消除噪声

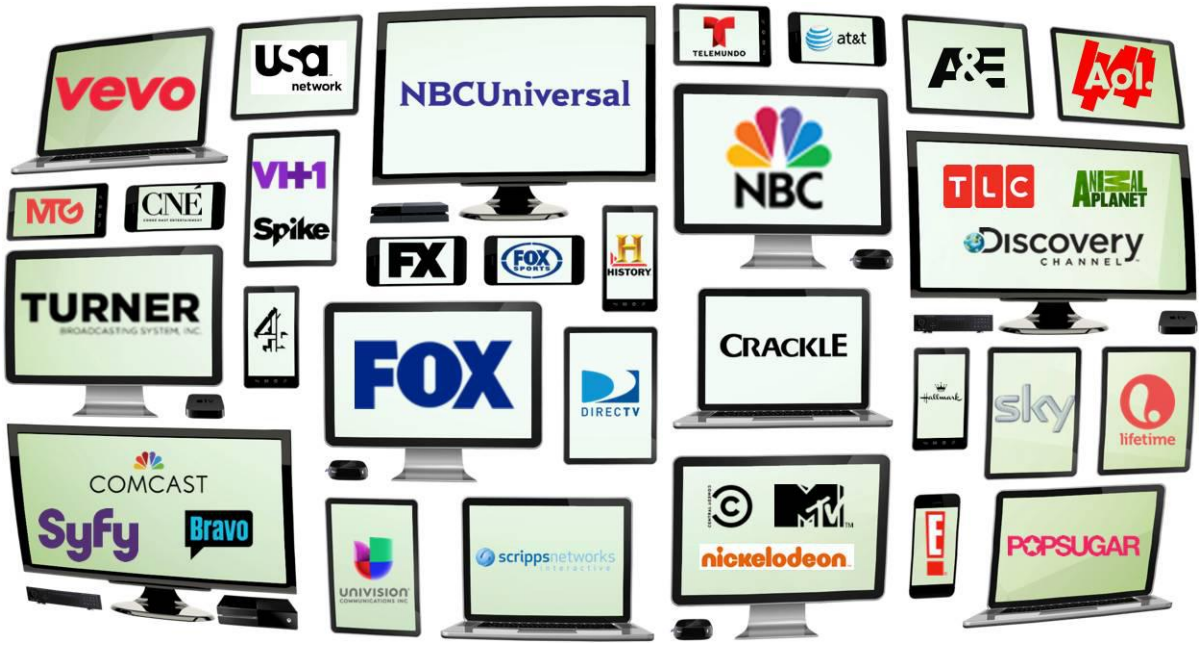


未来的想法

- 视频聚类
- 属性嵌入
- EM方法

关于FreeWheel

- 视频广告解决方案
  - 视频广告管理、投放、监测、预测、增值等业务
  - 支撑美国在线视频广告30%流量



FreeWheel

- 童有军
- Lead Engineer
- yjtong@freewheel.tv

Geekbang 极客邦科技 InfoQ

THANKS!