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　IT168.com　ChinaUnix　ITPUB

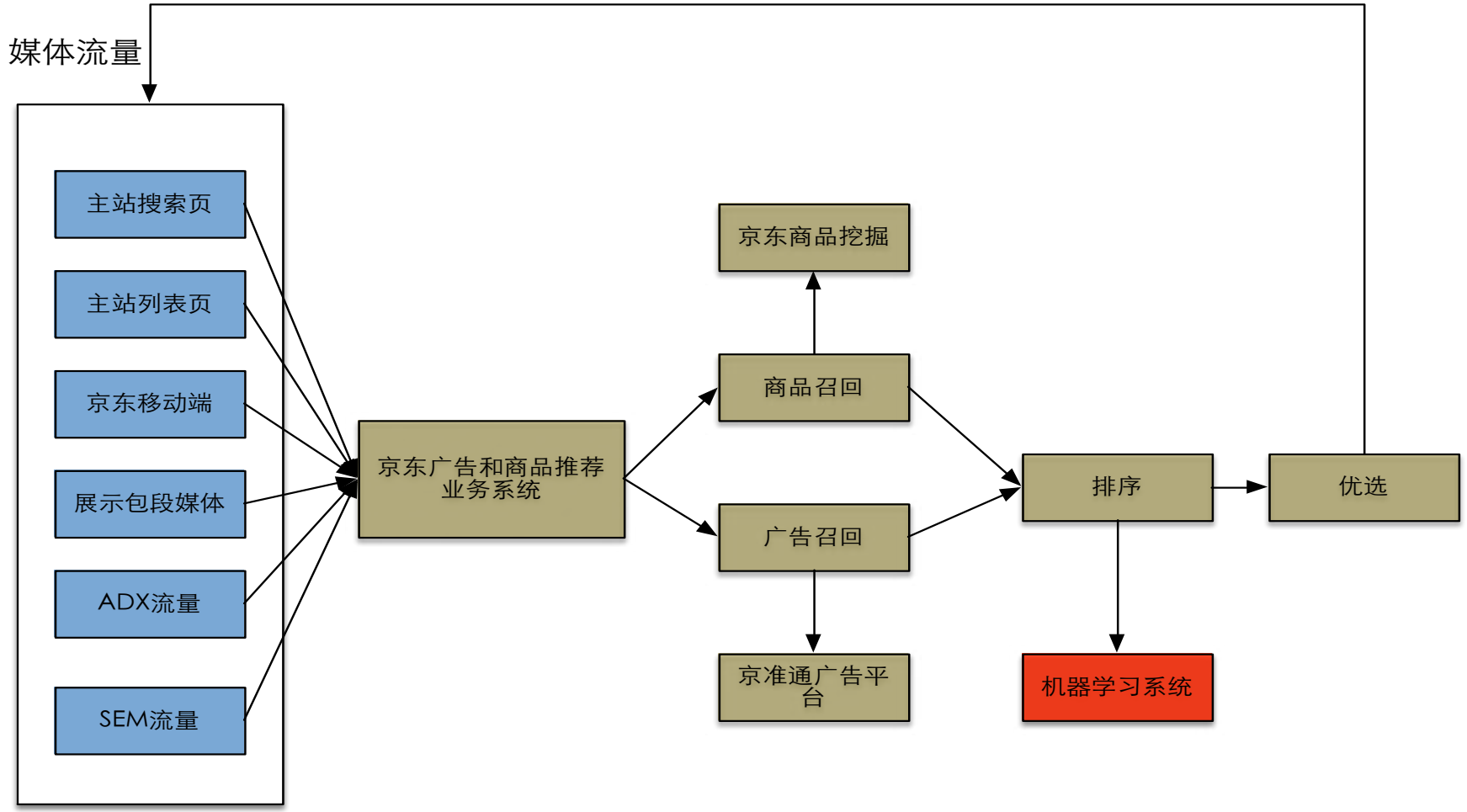
包勇军
京东

京东广告和推荐的机器学习系统实践

目录

- 背景介绍
- 浅层模型时代
- 深度学习时代

背景介绍|我们的业务


媒体流量
主站搜索页
主站列表页
京东移动端
展示包段媒体
ADX流量
SEM流量
京东广告和商品推荐业务系统
京东商品挖掘
商品召回
广告召回
京准通广告平台
排序
机器学习系统
优选

背景介绍|问题

- 主要解决的问题
  - 机器学习在排序算法中的应用
  - 特点：
    - 实时，在线
    - 广告，推荐的混合系统

目录

- 背景介绍
- **浅层模型时代**
- 深度学习时代

浅层模型时代|机器学习系统核心问题

- 模型算法
- 日志流
- 训练系统
- 特征系统
- 评估系统

浅层模型时代|模型算法

- 浅层模型算法：
  - 大规模稀疏性特征建模，lr
  - 核心优化方向：特征
    - 手工特征工程
    - 特征组合算法：
      - Fm/ffm
      - gbdt+lr

浅层模型时代|模型算法

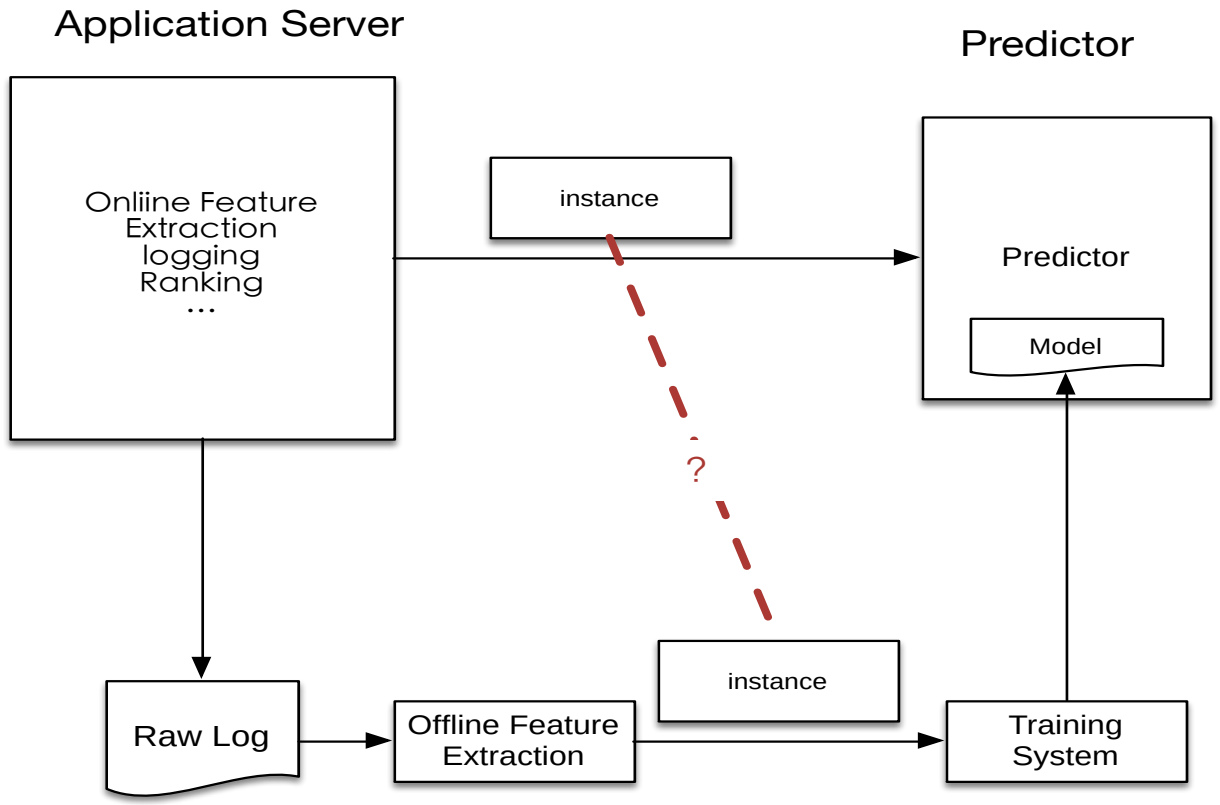
- Fm/ffm
  - 通过因式分解，减少数据稀疏性，有效学习特征组合
  - 参数规模：n^2降为k*n(k<<n, k为factor大小，n特征数目)
  - 问题：全组合的话，模型size = n*k，收益和资源的取舍

浅层模型时代|特征系统

- 特征系统主要问题：
  - 线上线下特征一致性
  - 根据经验，线上线下特征一致性的架构，在业务指标上能带来数量级的提升

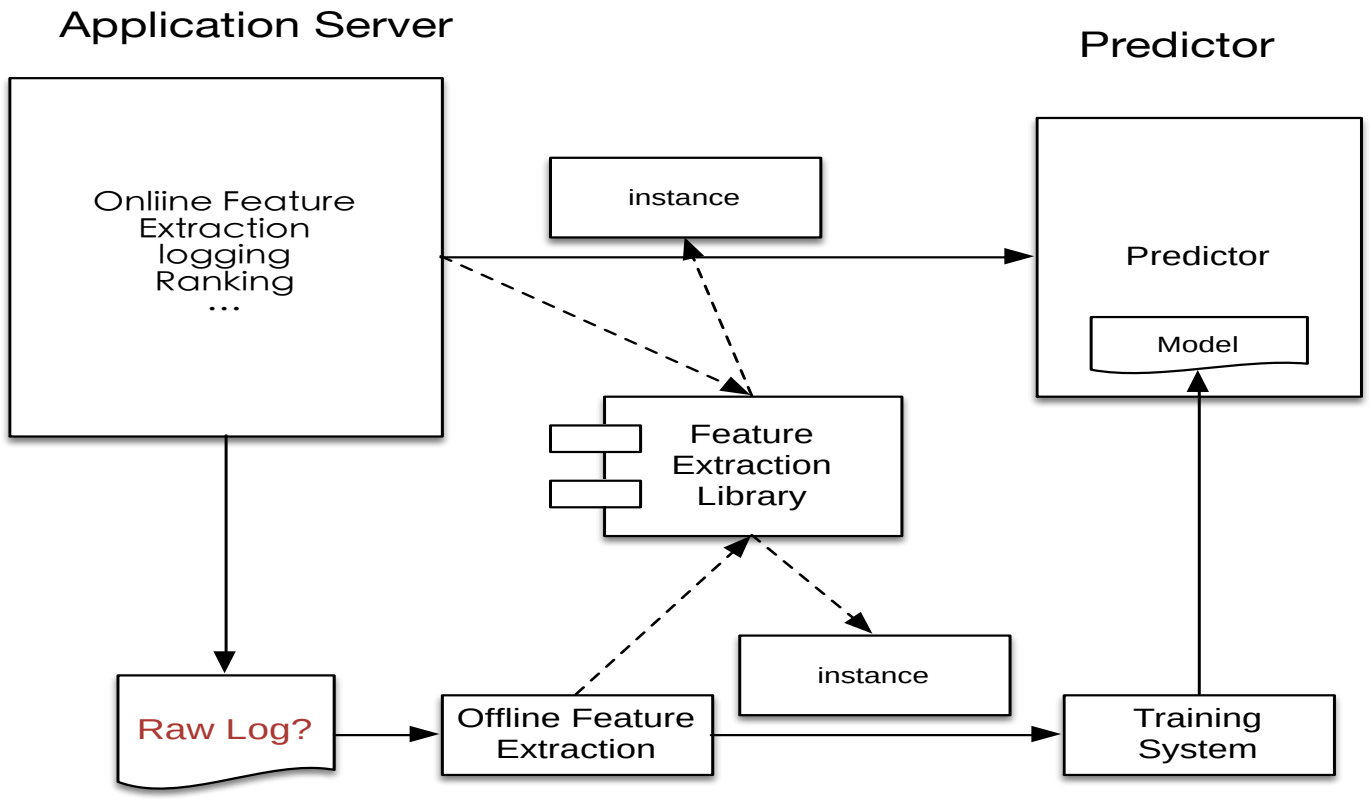
浅层模型时代|特征系统架构演化

- 第一版，开始引入机器学习模块，问题产生



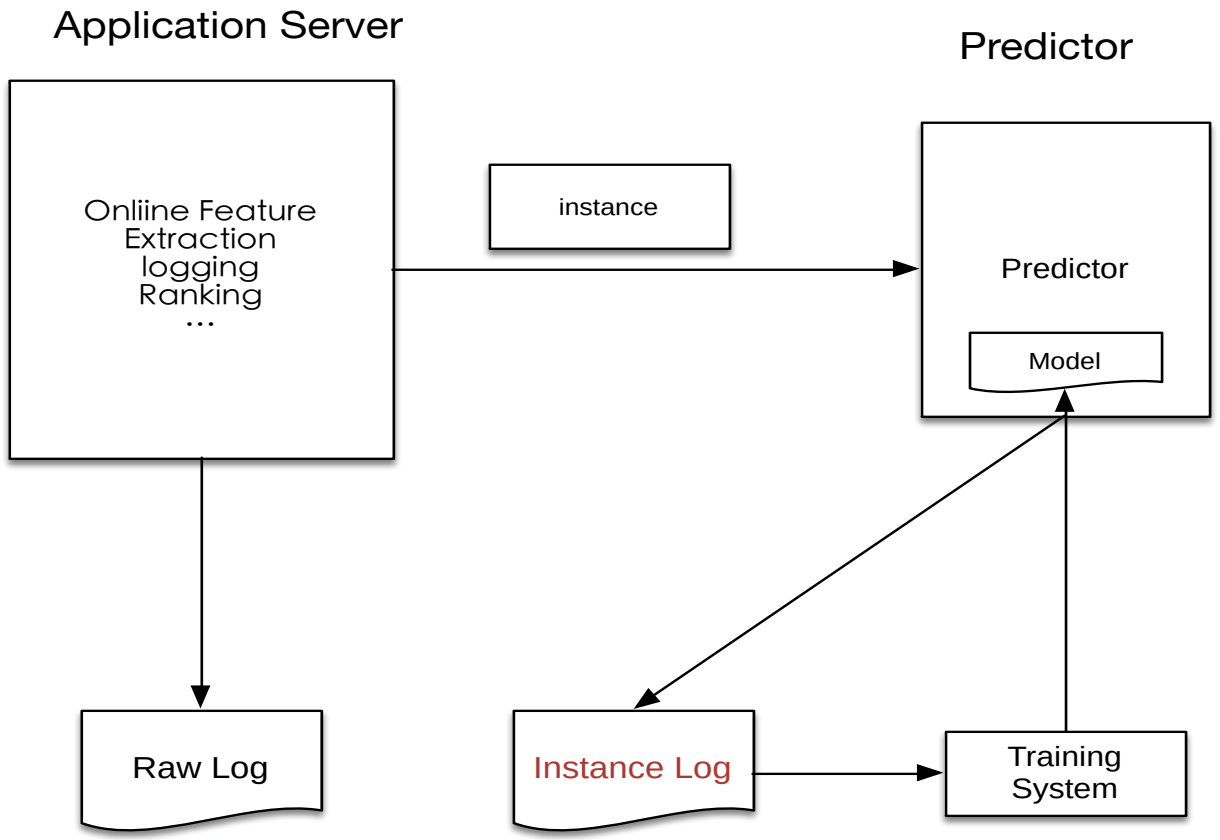
浅层模型时代|特征系统架构

- 第二版，解决代码不一致，代码复用



浅层模型时代|特征系统架构

- 第三版，解决数据不一致，彻底保证正确性



浅层模型时代|特征系统

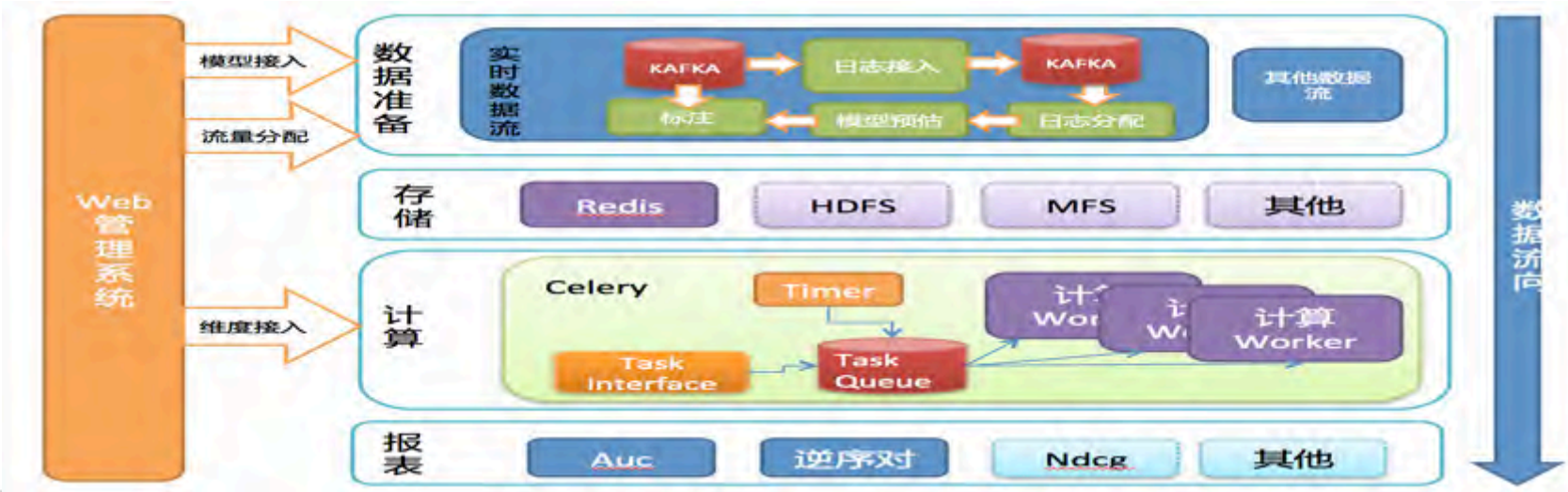
- 特征系统架构演变小结
  - 特征=数据源+抽取算法
  - 第一版是自然的选择
    - 机器学习系统是优化阶段的工作，先有日志后有机器学习
  - 第二版是策略效率为先的选择
    - 策略人员驱动后续的技术升级，离线代码驱动，先有离线代码后有在线代码
    - 日志量Double引发的资源担忧
    - 特征优化可以回朔历史数据，周期短
  - 第三版是保证策略收益的选择
    - 在线系统驱动特征升级，牺牲开发效率，保证正确性

浅层模型时代|模型效果评估

- 评估指标
  - AUC
  - Inverse Ratio
- 评估系统的主要问题：
  - 各种乌龙，结论不可信
  - 旧方案：离线工具评估离线指标
  - 新方案：在线系统评估离线指标

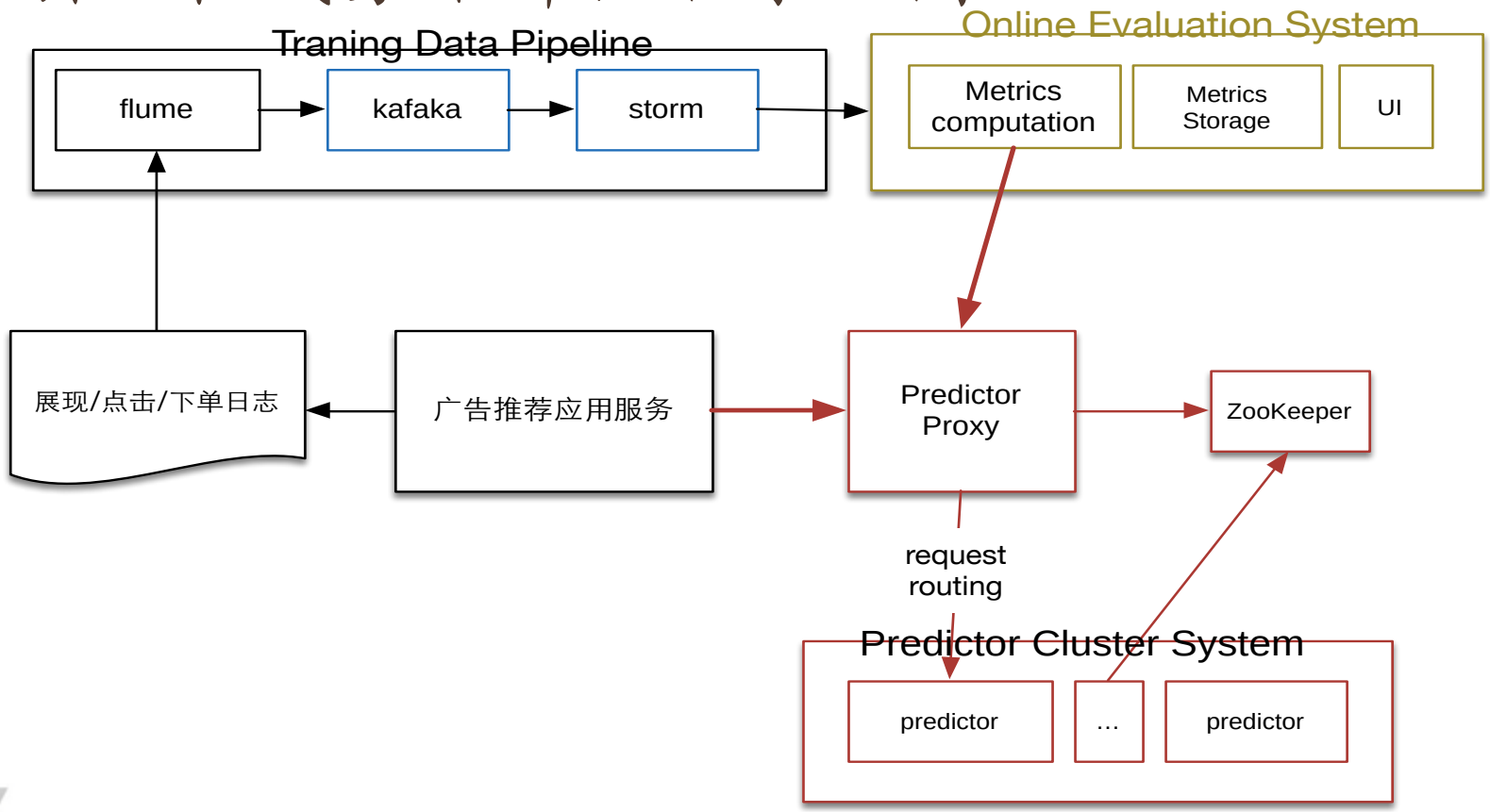
浅层模型时代|模型效果评估

- 在线旁路评估系统
  - 将在线predictor作为离线评估的inference工具
  - 将在线日志流作为离线评估数据
  - 离线测试模型接入在线predictor集群



浅层模型时代|旁路评估架构图

- 引入在线旁路评估后系统图

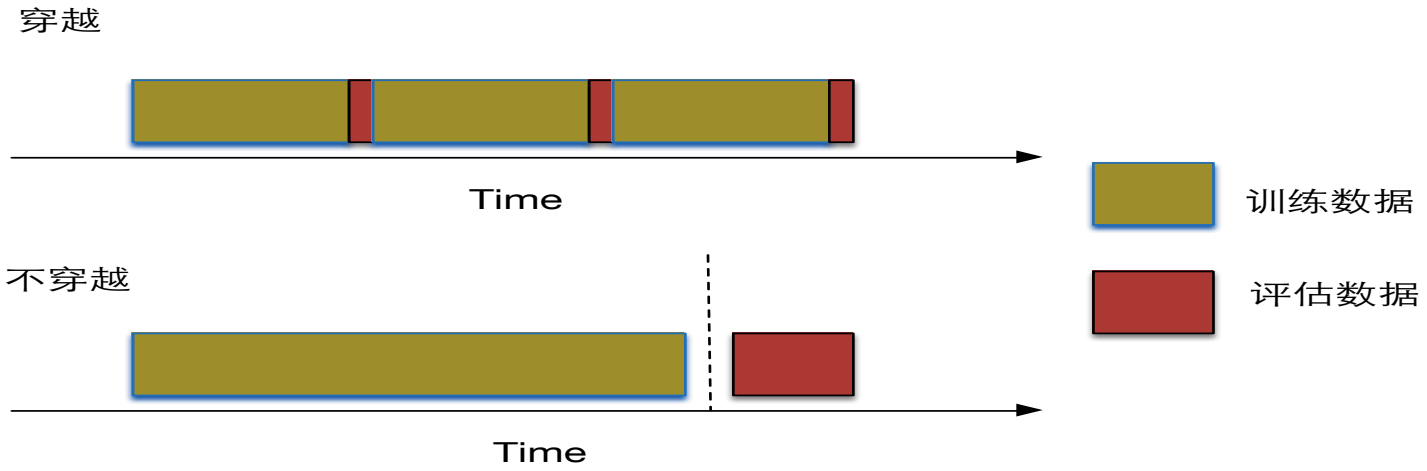


浅层模型时代|在线旁路评估

- 收益
  - 数据可比，可信
    - 工具到服务平台的升级
    - 避免数据diff和工具bug的干扰
    - 彻底解决在线实时服务模型中的评估穿越问题

浅层模型时代|模型效果评估

- 在线实时服务模型中的评估穿越问题
  - Unseen data，历史数据预估新数据
  - 数据分布变化更快，泛化性要求更高
    - 推荐中的新兴趣点
    - 广告中的新广告



浅层模型时代|训练系统

- 浅层模型训练系统的核心问题： 大数据的效率问题
  - Sampling
  - Distributed training， libfm on vowpal wabbit
  - Incremental
  - Online learning:
    - Assumption: stationary -> concept drift
    - 好处：
      - state track， 时效性
    - 问题：
      - 系统复杂，需要增加实时计算系统
      - 更新频繁，增加了系统耦合
      - 特征和算法升级麻烦

浅层模型时代|多目标优化

- 业务目标：广告收入（year 2014）
  - eCpm = pCtr * bid
  - pCtr：通过机器学习进行点击率预估

浅层模型时代|多目标优化

- 多目标优化
  - 广告收入+GMV (year 2015)
  - RankingFunction=pCtr1*(a*pValue+b*pGmv)
  - 三个模型： pCtr, pGmv, pValue
- 多模型方案的问题
  - 分目标优化，策略升级不能同步
  - 点击后模型Gmv的训练数据稀疏

浅层模型时代|多目标优化方案

- 多目标优化
  - One model方案
    - 收入 + Gmv一起建模，策略同步，数据更丰富
    - Pairwise + Pointwise
      - Combine regression and rank

$$\min_{\mathbf{w}\in\mathbb{R}^m} \alpha L(\mathbf{w}, D) + (1-\alpha) L(\mathbf{w}, P) + \frac{\lambda}{2}||\mathbf{w}||_2^2$$

      - L(w, D) regression loss, L(w, P) pairwise rank loss
      - Rank loss保证不同label的序关系，在rare events场景，能提升regression的效果
      - Regression loss拟合绝对值，保持分布稳定，用于广告的二价计费

目录

- 背景介绍
- 浅层模型时代
- 深度学习时代

深度学习时代

- 为什么引入深度学习？
  - 非线性模型
    - LR是通过各种特征组合来实现
      - 人工特征组合，高维线性模型建模非线性
      - Libfm, depth 2
      - GBDT+LR, depth 3
      - 大数据背景下，DNN更通用
  - 优化方式算法驱动
    - Manual feature engineering->Feature Learning
      - 浅层模型： Raw data->hand craft->feature;
      - 深度学习： Raw data->algorithm->feature;

深度学习时代|面临的问题

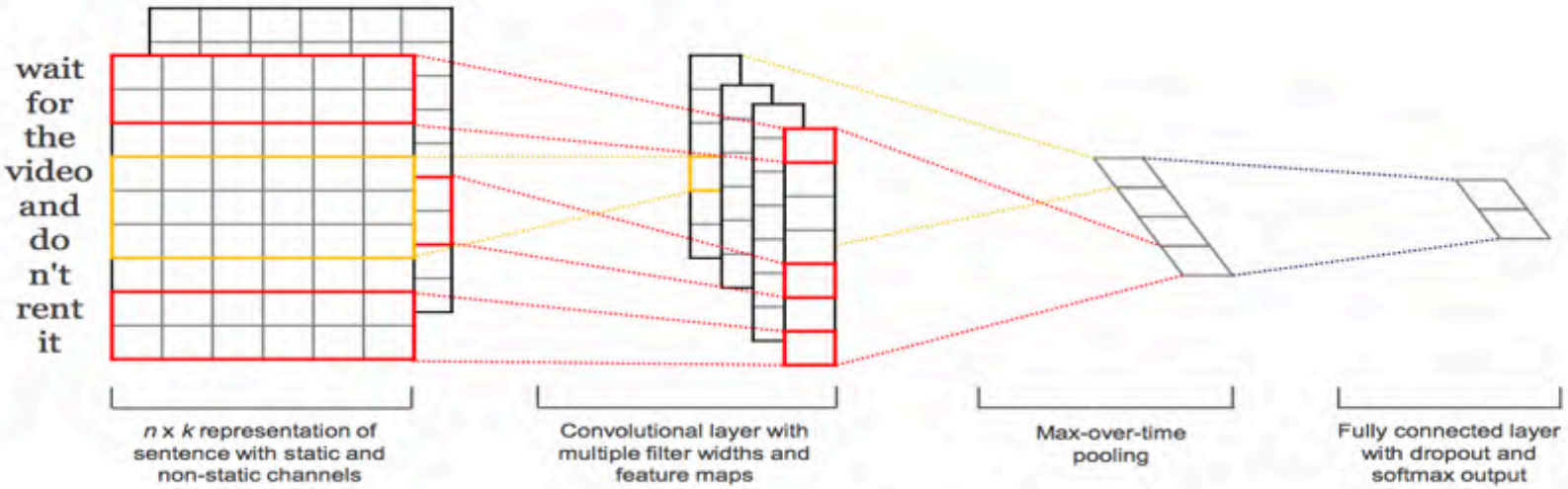
- 引入深度学习面临的问题
  - 现有算法系统以及效果如何平滑过渡
  - 离散特征如何建模
    - billion级别，海量，稀疏

深度学习时代|建模选型

- 稀疏离散特征的DNN建模方法：
  - 离散特征数值化：把特征离散值映射到连续型的数值空间
    - Embedding法
      - 每一个样本都是几亿维
    - 稀疏样本转稠密向量表示
  - CNN方法：文本转图像

深度学习时代|建模选型

- CNN法|样本文本当成图像
  - 1-of-n encoding
  - Input embedding

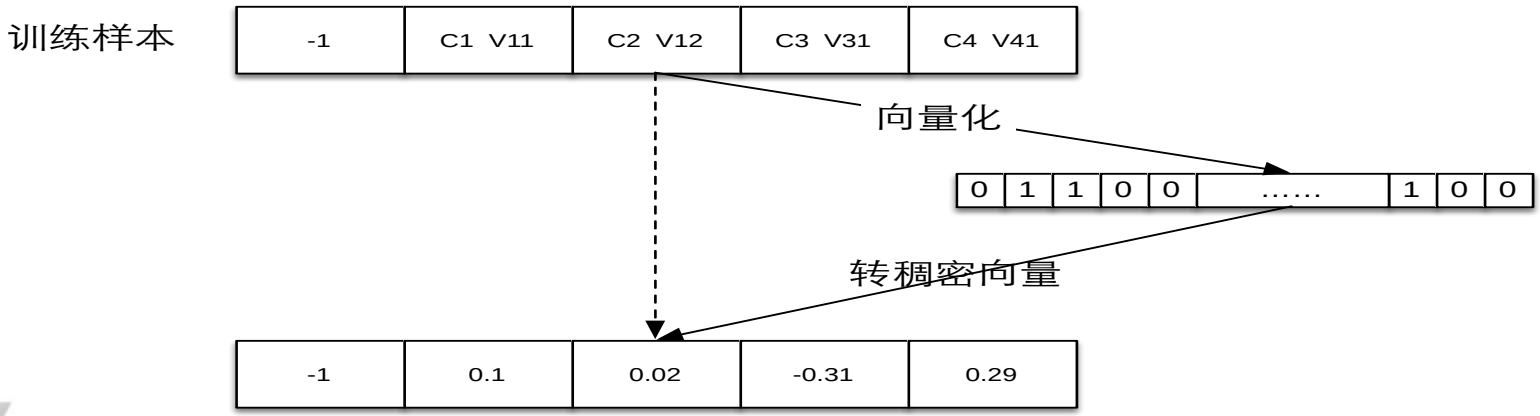


深度学习时代|建模选型

- CNN法
  - 效果：AUC 有明显提升
  - 问题：
    - 10倍的在线预估cost，在线架构的大量优化工作
    - 消耗资源大，性价比低

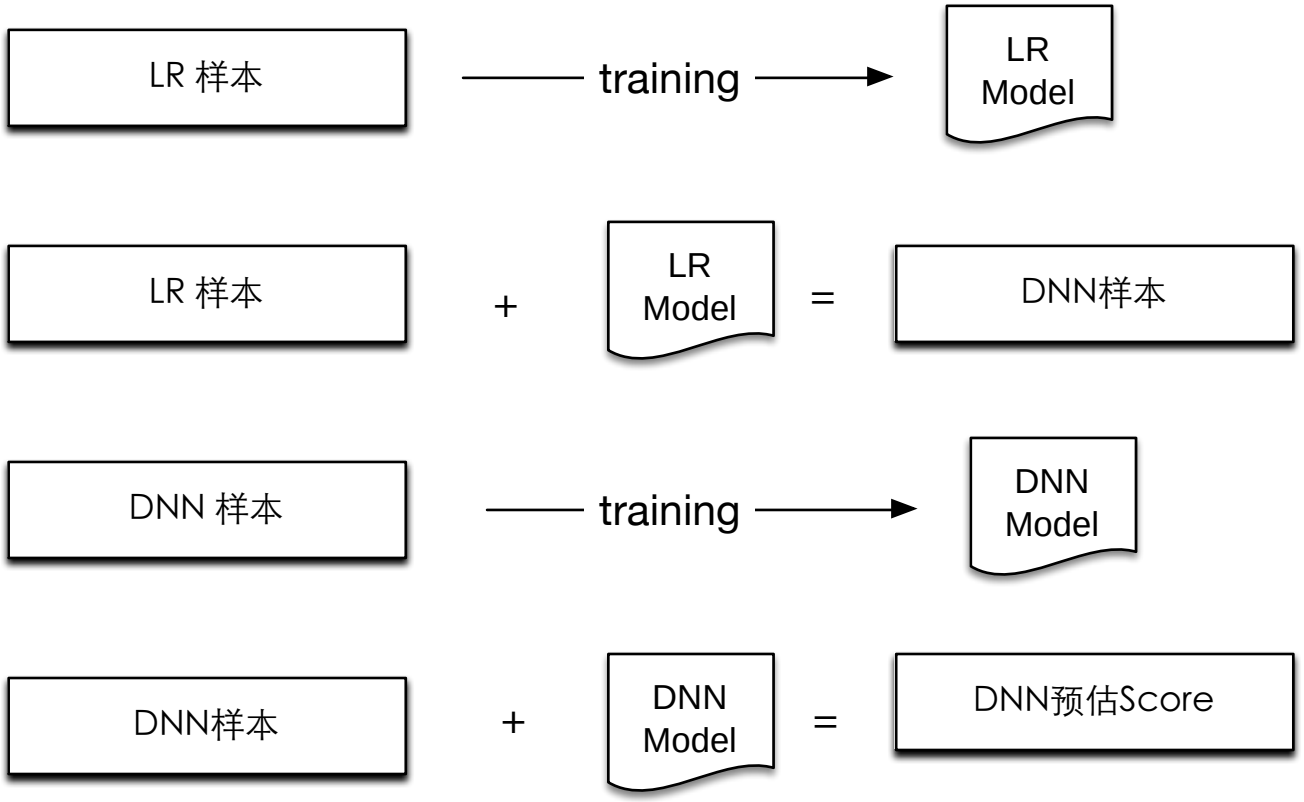
深度学习时代|样本表示

- 稀疏样本转稠密表示
  - 每一个特征类是输入的一维
  - 离散特征值映射到连续值
    - 后验点击率
    - LR weight



深度学习时代|系统方案

- LR model->DNN法, LR to Dnn

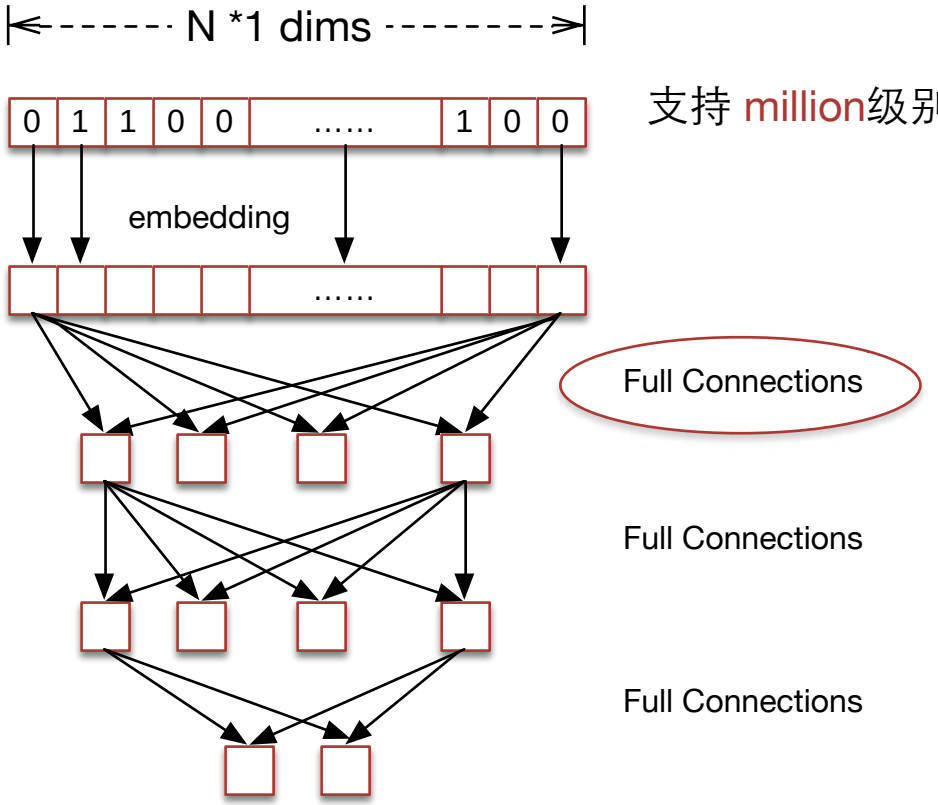


深度学习时代|方案总结

- LR to DNN方法小结
  - 效果：对比libfm模型，AUC +2%
  - 问题：
    - LR权重不稳定，DNN层效果波动
    - 系统复杂，增加特征周期长，升级困难

深度学习时代|方案升级

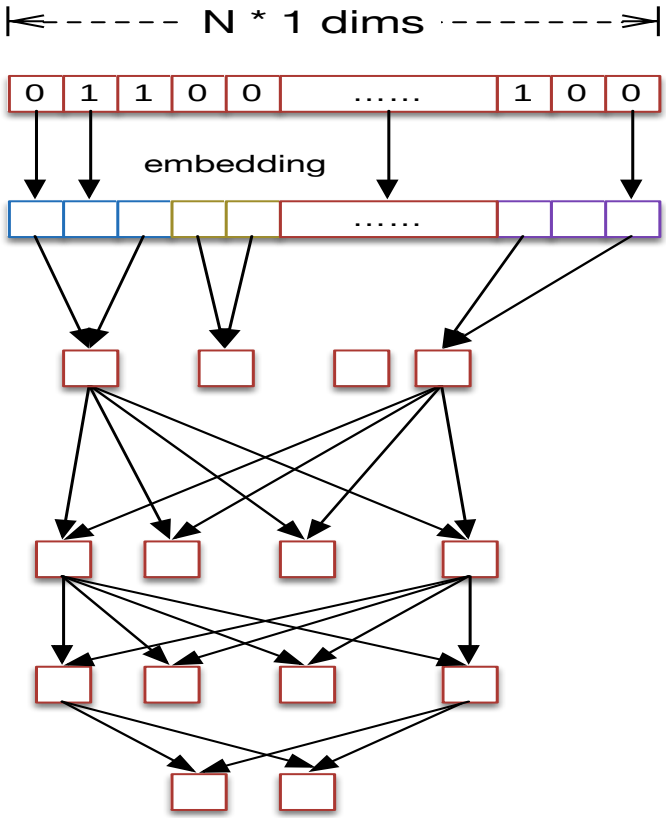
- DenseDNN with LR embedding, 合并LR和DNN到一个网络



支持 million级别

深度学习时代|方案升级

- SparseDNN with LR embedding, pooling思想降低参数规模



支持billion级别

pooling

k dims, k<<N

Full Connections

Full Connections

Full Connections

深度学习时代|升级后效果

- SparseDNN with LR embedding方法，效果:
  - 对比Lr to Dnn，AUC累计提升2%-3%
  - 无权重波动，系统稳定;
  - Training together，无各种穿越问题;
  - One model统一结构，系统简单,更易继续优化扩展
    - 离散特征，LR embedding 接入
    - 连续特征，图像CNN embedding，行为RNN embedding 接入

深度学习时代|DNN训练系统

- 现有开源框架问题
  - 10亿特征，150亿的样本
  - 现有开源框架的问题
    - Theano, Caffe, mxnet, Petuum, DMTK, Tensorflow
    - 多机支持，GPU不能解决IO负载大的问题
    - AllReduce方案，模型全量同步，通信开销大
    - 稀疏性的支持，大规模稀疏矩阵运算

深度学习时代|DNN训练系统

- 基于开源自研的分布式训练系统
  - Theano + ParameterServer架构
    - 尽量复用现有开源框架
    - 深度定制Theano，以支持大规模稀疏矩阵运算
    - ParameterServer作为参数交换的机制
    - Downpour SGD实现
- 系统性能
  - 10亿稀疏特征+5层神经网络，150亿样本，4小时训练

THANKS

SequeMedia 盛拓傳媒 IT168.com ChinaUnix ITPUB