1.简介

1.1 技术架构

f(用户信息，资源信息，上下文信息，用户行为序列) = 推荐结果

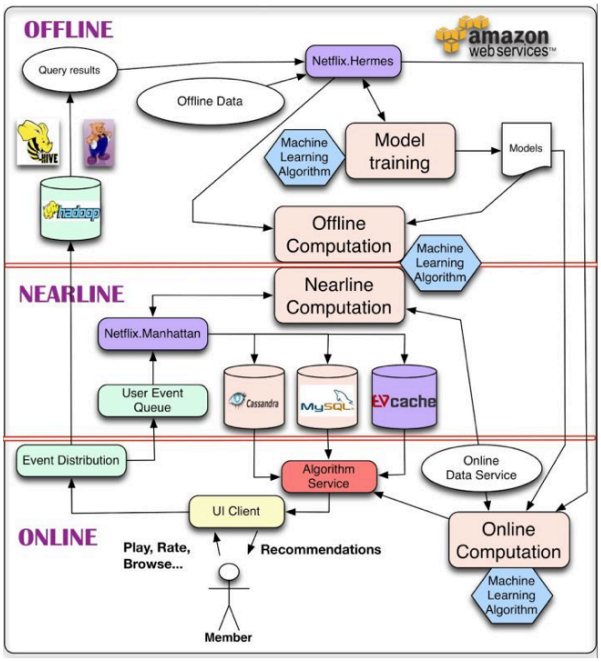
1.2 搜广推

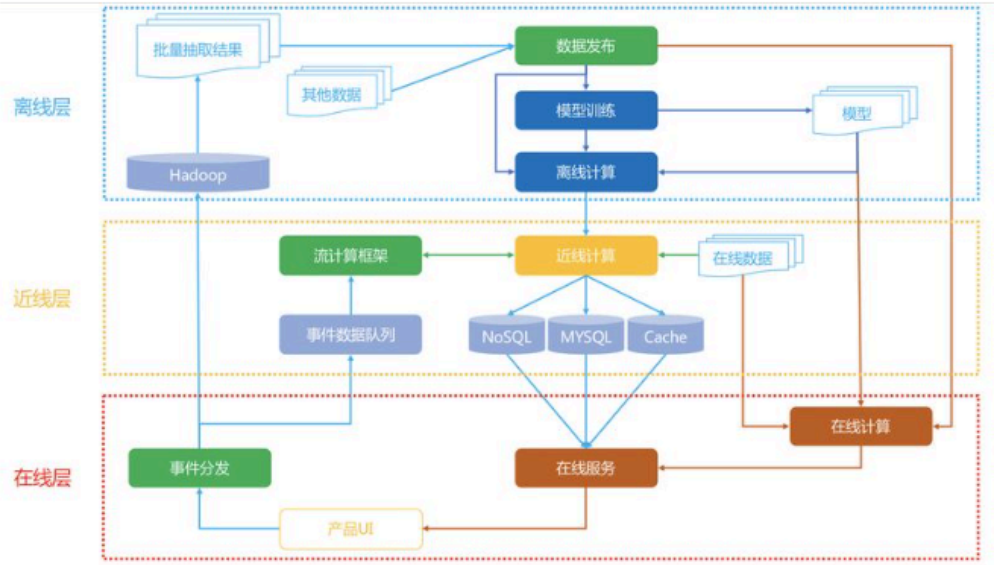
- 搜索：围绕着搜索词的信息高效获取问题
- 广告：直接增加公司收入
- 推荐：提高用户留存和活跃度
  具体区别-知乎-王喆

1.3 架构

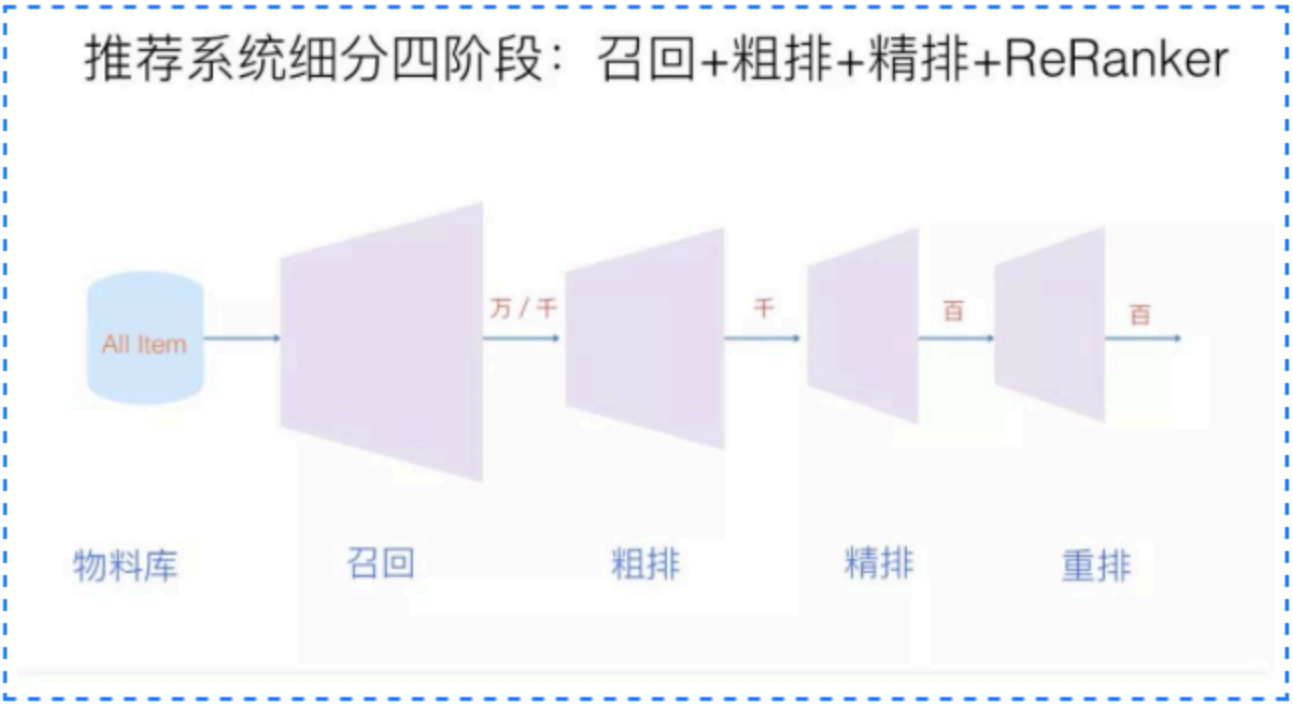
1.3.1 系统架构

1. 离线层：不用实时数据，不提供实时响应；(大量)
   可以每天更新一次。
2. 近线层：使用实时数据，不保证实时响应；(几分钟)
   可以在用户访问时更新。
3. 在线层：使用实时数据，保证实时在线服务。(几十毫秒)
   在用户访问时实时响应，如开屏时的推荐。
   Netflix(2013)


OFFLINE
amazon web services
Query results
Netflix.Hermes
Offline Data
Machine Learning Algorithm
Model training
Models
Offline Computation
Machine Learning Algorithm
NEARLINE
Nearline Computation
Netflix.Manhattan
User Event Queue
Cassandra
MySQL
EVcache
Event Distribution
Algorithm Service
Online Data Service
UI Client
ONLINE
Play, Rate, Browse...
Recommendations
Online Computation
Machine Learning Algorithm
Member


离线层
批量抽取结果
其他数据
数据发布
模型训练
模型
Hadoop
离线计算
近线层
流计算框架
近线计算
在线数据
事件数据队列
NoSQL
MYSQL
Cache
在线计算
在线层
事件分发
在线服务
产品UI

1.3.2 算法架构



- **召回**：不需要十分准确，但不可遗漏。快速稳定，兴趣多元、内容多样。
  - 非个性化召回、个性化召回
    - 个性化召回：content-based、behavior-based、feature-based
- **粗排**：兼顾精准性和低延迟。
- **精排**：精准性优先，目标ctr、cvr(点击率、转化率)，指标AUC。
  - 样本、特征、模型。
- **重排**：多样性优先，目标Point Wise、Pair Wise、List Wise(点对、对对、列表)，指标NDCG。
  - 基于运营策略、多样性、context上下文。

1. **画像层**

- 文本理解
  - RNN、TextCNN、FastText、Bert
- 关键词标签
  - TF-IDF、Bert、LSTM-CR
- 内容理解
  - TSN、RetinaFace、PSENet

2. **召回/粗排**

- 经典模型召回
  对user和item分别打上Embedding，然后user与item在线进行KNN计算实时查询最近领结果作为召回结果

    - FM、双塔DSSM、Multi-View DNN
- 序列模型召回
  有监督Next Item Prediction，无监督Sum Pooling
    - CBOW、Skip-Gram、GRU、Bert
- 用户序列拆分
  把用户行为序列打到多个embedding上(类似聚类)
    - Multi-Interest Network with Dynamic Routing for Recommendation at Tmall
- 知识图谱
  可解释性好但是效果差
    - KGAT、RippleNet
- 图模型
  编码的是静态知识，而不是用户比较直接的行为数据，所以效果一般
    - GraphSAGE、PinSage

3. **精排**

- 特征交叉模型
    - DCN、DeepFM、xDeepFM
- 序列模型
  关注用户此刻的兴趣向量（user interest vector）
    - DIN、DSIN、DIEN、SIM
- 多模态信息融合
    - Image Matters: Visually modeling user behaviors using Advanced Model Server、UMPR
- 多任务模型
  点击率模型(标题党)、时长模型(长视频)、完播率模型(短视频)
    - ESSM、MMoE、DUP
- 强化学习
    - DQN、Reinforcement Learning for Slate-based Recommender Systems: A Tractable Decomposition and Practical Methodology
- 跨域推荐
  比如QQ音乐的信息如何使用在腾讯视频上
    - DTCDR、MV-DNN、EMCDR

电子书-FunRec
纸质书-推荐系统技术原理与实践