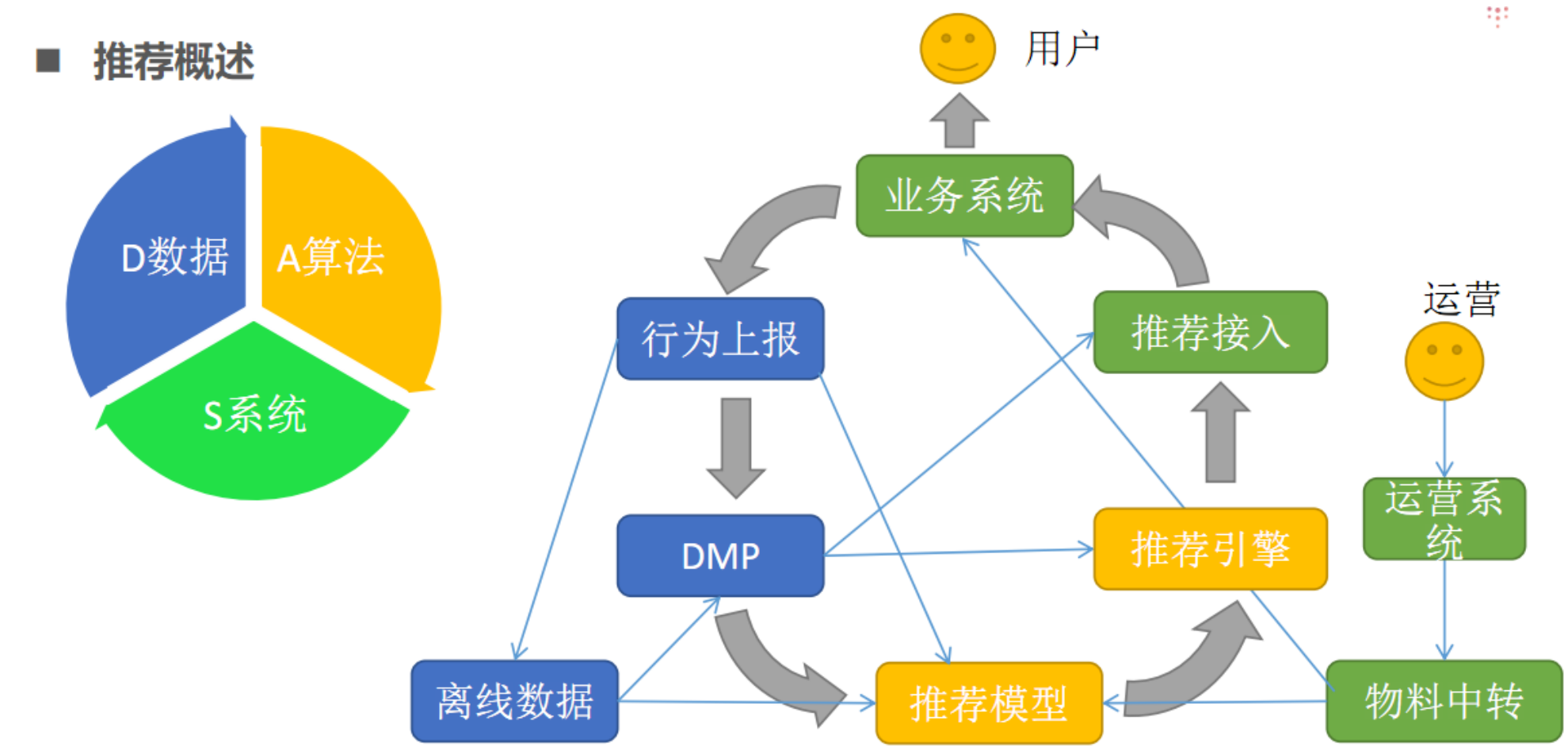
京东推荐

xpguo

Reference: 《微信购物（京东）个性化推荐实战》--马兴国

ABC赋能

ABC=(AI + BigData + Cloud)



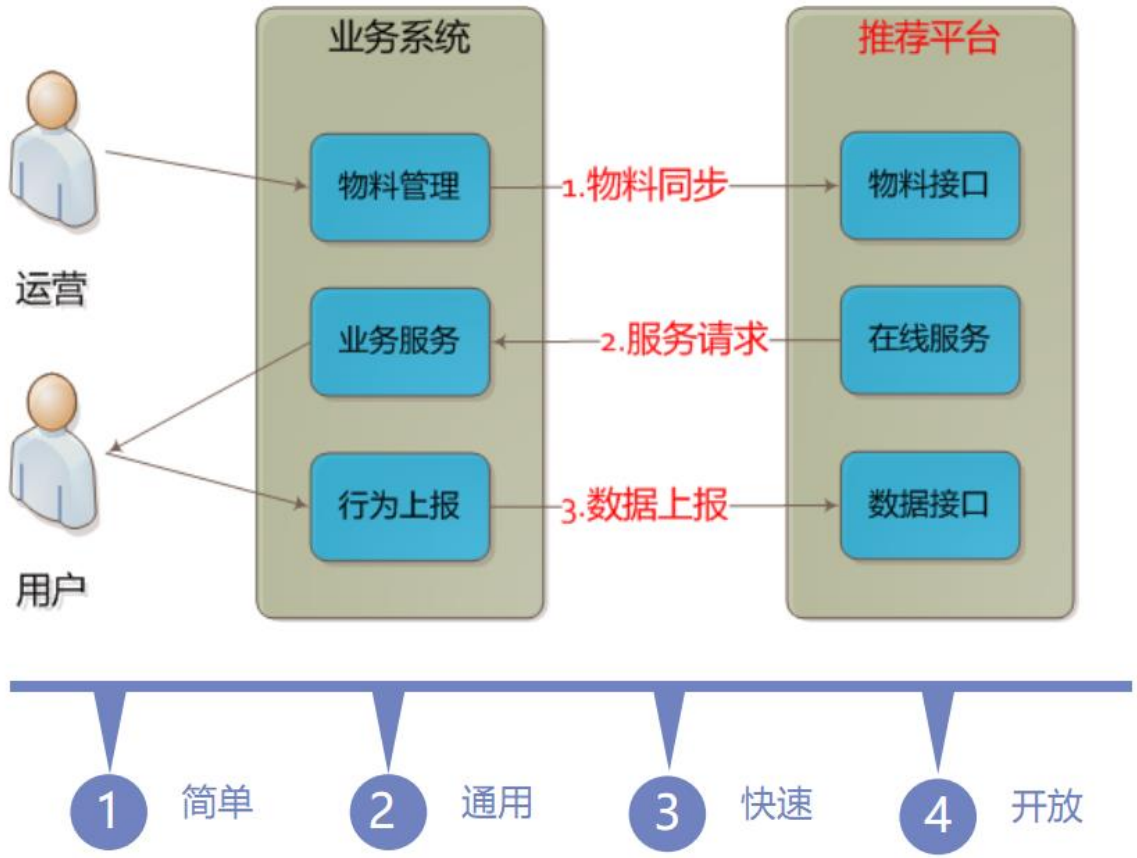
平台架构


业务层
业务服务
词推荐
素材推荐
商品推荐
资讯推荐
定向投放
其他
接入层
推荐接入
物料同步
统一请求
行为上报
智能分流
第三方引擎
TX:TEG
TX:WXG
TX:SNG
JD推荐
JD广告
引擎层
推荐引擎
召回策略
打分排序
重排策略
计算层
推荐模型
实时训练
离线训练
模型融合
模型评估
数据层
DMP(大数据管理平台)
查询服务
用户画像
物料画像
数据分析
毕加索用户画像
基础平台
大数据平台
SPARK
HADOOP
CM(Cloudera Manager)
Sklearn/TensorFlow
京东大数据平台
数据高速通路
FLume
KAFKA
支撑系统
运营系统
配置系统
调度系统
监控系统
评估系统

平台接入

■ **平台接入**

- 物料同步
  - 统一格式
  - 增量
  - 全量
- 服务请求
  - 标准化
  - 高性能
  - 智能化
- 行为上报
  - 统一格式
  - 实时
  - 离线



推荐引擎

- 推荐引擎

召回策略

- 搜索召回
- 规则召回
- 行为召回
- 个性化特征召回
- 关联规则前键召回
- 热度召回
- 冷启动召回
- 其他

排序打分

- 热度打分
- 线性打分
- 树形打分
- 离线打分
- 实时打分
- 算法融合
- 粗排精排
- 其他

重排策略

- EE策略
- 特征打散策略(最小距离，多路归并，多特征)
- 已购降权(过滤，沉底)
- 特征加权
- 其他

数据处理

■ 数据处理

数据是基础

收集数据
- 选择需要的数据源。
- 数据量要够，样本要完备。
- 收集方式：拉：爬虫；推：上报等。

分析数据
- 内容：目标分布，特征分布，目标特征关系，特征间关系，完整性等。
- 方式：离线，实时，融合。
- 工具：Excle, Shell(awk), Python, R,Mysql, Hadoop, Spark, Matlab等。

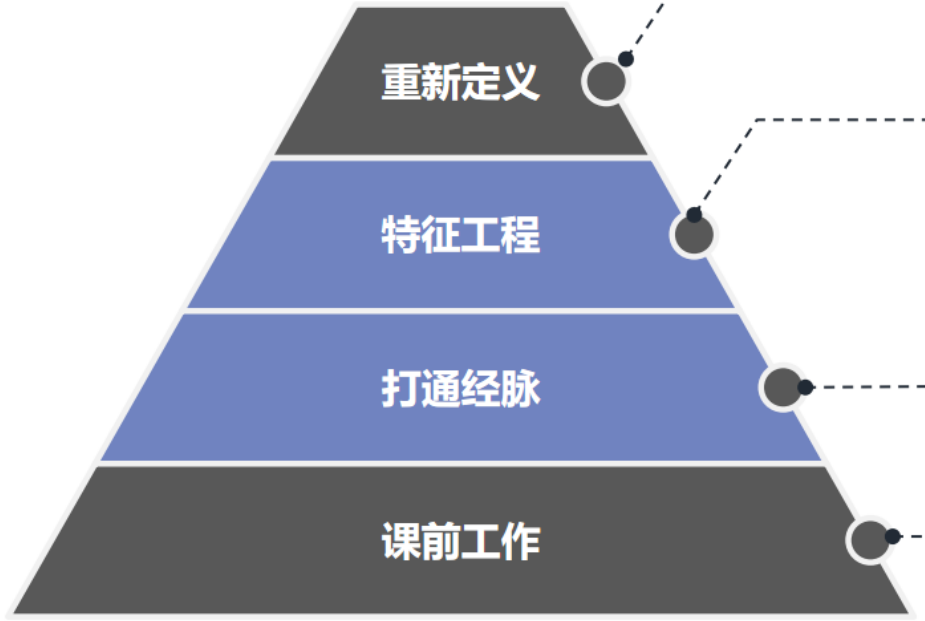
清洗数据
- 系统脏数据：非业务正常请求的系统外脏数据，刷请求，爬虫请求，刺探请求。
- 业务脏数据：假曝光数据，前端预加载请求等。最后一个点击之前的曝光作为有效曝光。
- 目标外数据：和数据处理目标不相干数据过滤。

格式数据
- 数据变换：归一化，离散化，OneHot，Scale分级(年龄，价格区间等)，特征关联等
- 采样：下采样(TraceId关联，简单按比例随机抽样等)；上采样(1/CTR，倍数等)
- 稀疏处理：其他 (常用)，默认值，最热值(离散型)，均值(正态分布类连续型)，特征剔除(80%无数据)等。

机器学习军规

■ 机器学习军规

ML在实际工作更多的是工程问题，其次才是算法问题。
优先从工程中要效果，当把这部分榨干后，再考虑算法的升级。



重新定义
Slowed Growth, Optimization Refinement,
当效果进入稳定期，寻找本质上新的信息源，而不是优化已有的效果稳定的特性

特征工程
Feature Engineering
优先用直观的特征，而不是学习出来的特征；在错误和BadCase中发现突破点；保证服务与训练一致性的最好方法是将服务时的特征保存下来，用到训练过程中去。

打通经脉
First Pipelin
第一个模型要简单，但是架构要好；数据处理，干净，完备，格式化；使用可解释性强的模型可降低debug难度。

课前准备
Before Machine Learning
在动手之前先设计和实现评价指标。
不要使用过于复杂的规则系统，使用机器学习系统。

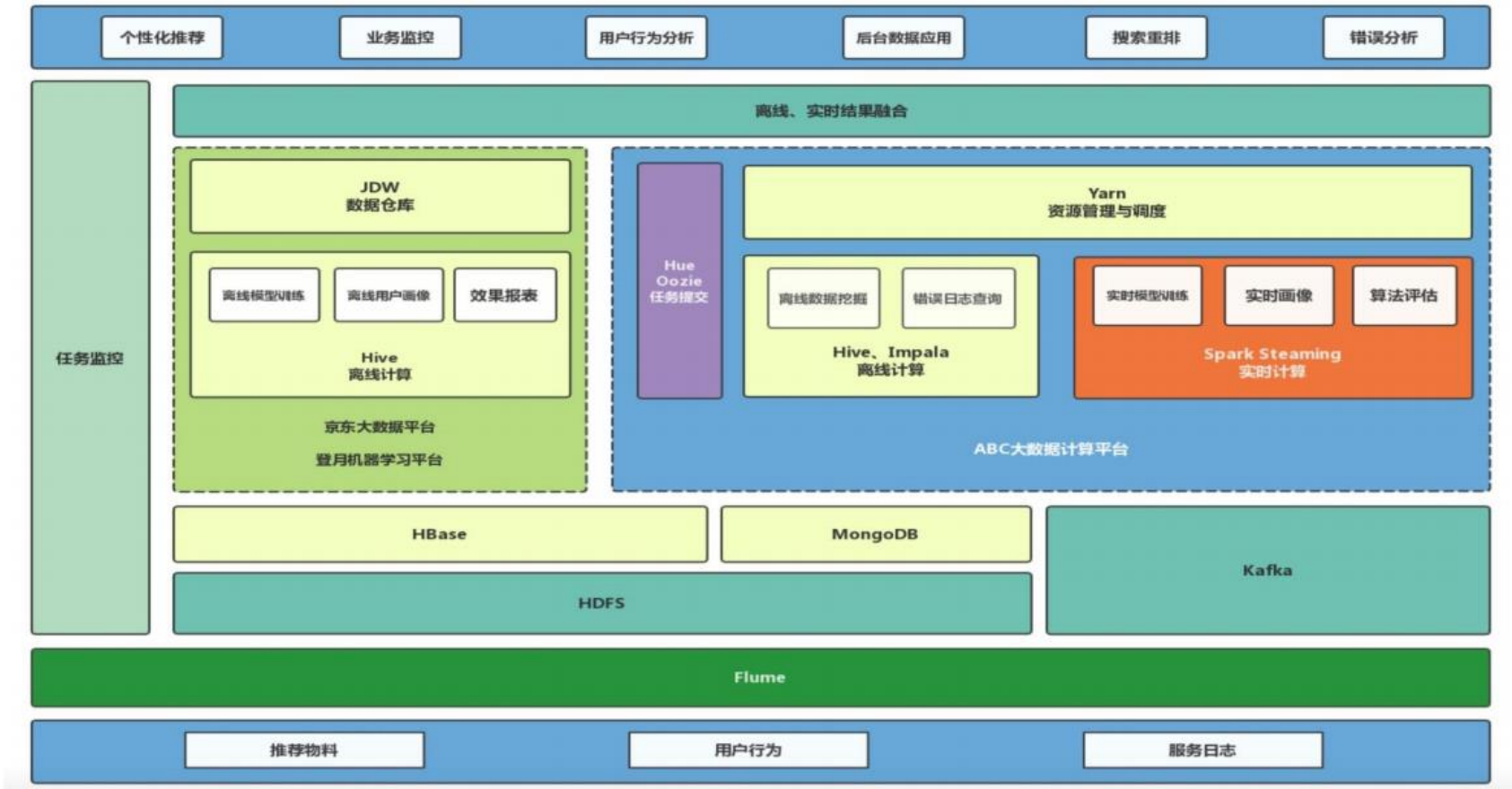
如果说大神做的是95分以上的系统，
只要我们对系统、数据做到足够的优化，也可以做出80分的系统。

【Rules of Machine Learning:Best Practices for ML Engineering】
http://martin.zinkevich.org/rules_of_ml/rules_of_ml.pdf

用户画像


业务应用
精准营销
广告投放
场景营销
用户包提取
其　他
个性化产品
商品推荐
素材推荐
资讯推荐
搜索个性化
数据分析
用户分析
业务分析
风控分析
其　他
WQ画像系统
对外接口
查询服务
查询API
查询Portal
数据存储
HDFS
REDIS
MYSQL
画像计算
实时计算
离线计算
数据融合
数据采集
数据分析
数据清洗
数据格式化
支撑系统
管理后台
监控系统
调度系统
配置系统
内外数据
离线数据
BDP
实时业务数据
搜索
浏览
加车
下单
其他
第三方数据
社交数据

大数据平台


个性化推荐
业务监控
用户行为分析
后台数据应用
搜索重排
错误分析
任务监控
离线、实时结果融合
JDW
数据仓库
离线模型训练
离线用户画像
效果报表
Hive
离线计算
京东大数据平台
登月机器学习平台
Hue
Oozie
任务提交
Yarn
资源管理与调度
离线数据挖掘
错误日志查询
Hive、Impala
离线计算
实时模型训练
实时画像
算法评估
Spark Steaming
实时计算
ABC大数据计算平台
HBase
MongoDB
Kafka
HDFS
Flume
推荐物料
用户行为
服务日志

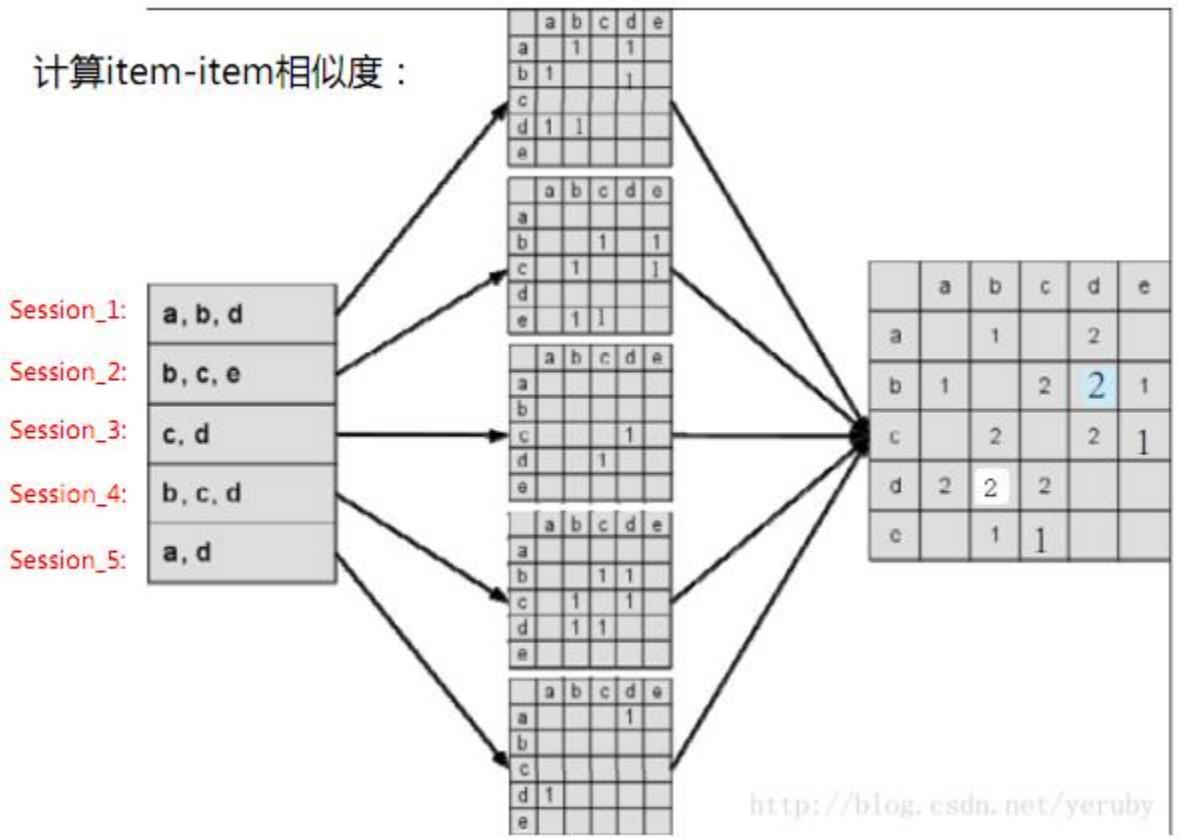
JD广告算法学习

xpguo

Reference: 《广告精准投放实践》--谢礼明

retargeting

Item-cf



i2i相似度：

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

做去热处理：

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

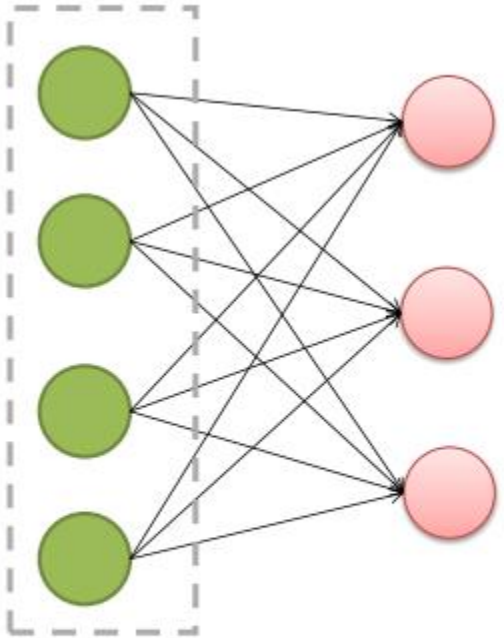
sku2vec

如何每个skuid不是字符串，而是一个向量。

[0.5,0.15,.....,0.06]

[0.5,0.20,.....,0.05]

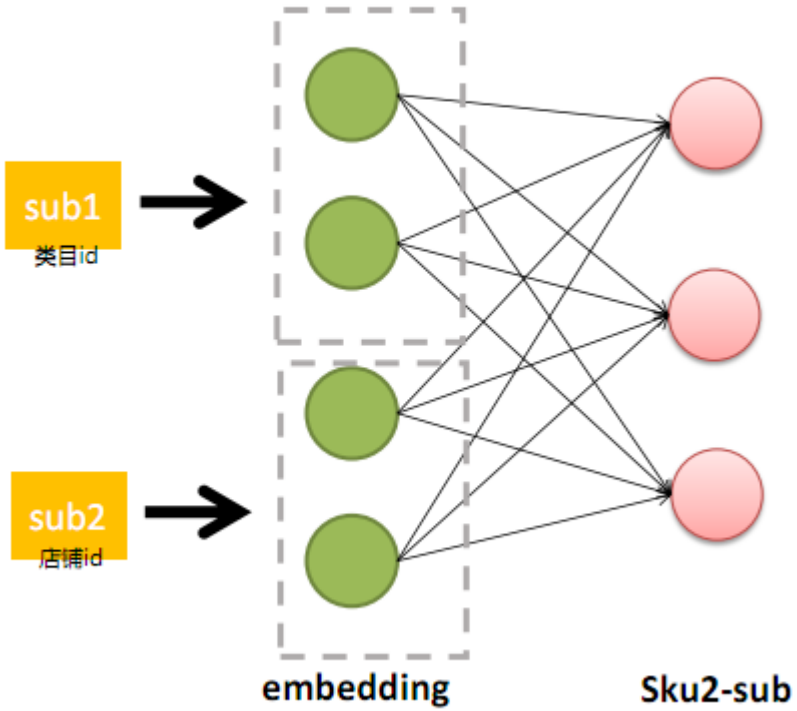
sku1



embedding

sku2

sku2vec



1，采用skuid->(类目id,店铺id,....)后，增强了类目,店铺的头部信息，在计算sku2sku 相似度时，同一类目，店铺的sku更靠前。

2，**并行训练**。由于sku库及样本特别大，我们将sku划归大类，分别进行并行训练。加快训练速度。

3，采用tensorflow训练。

LR -> GBDT+LR -> Wide_n_Deep -> DCN

下一步规划，

通过DCN解决feture -cross

Cross layer定义：

$$x_{l+1} = x_0 x_l^T w_l + b_l + x_l = f(x_l, w_l, b_l) + x_l,$$

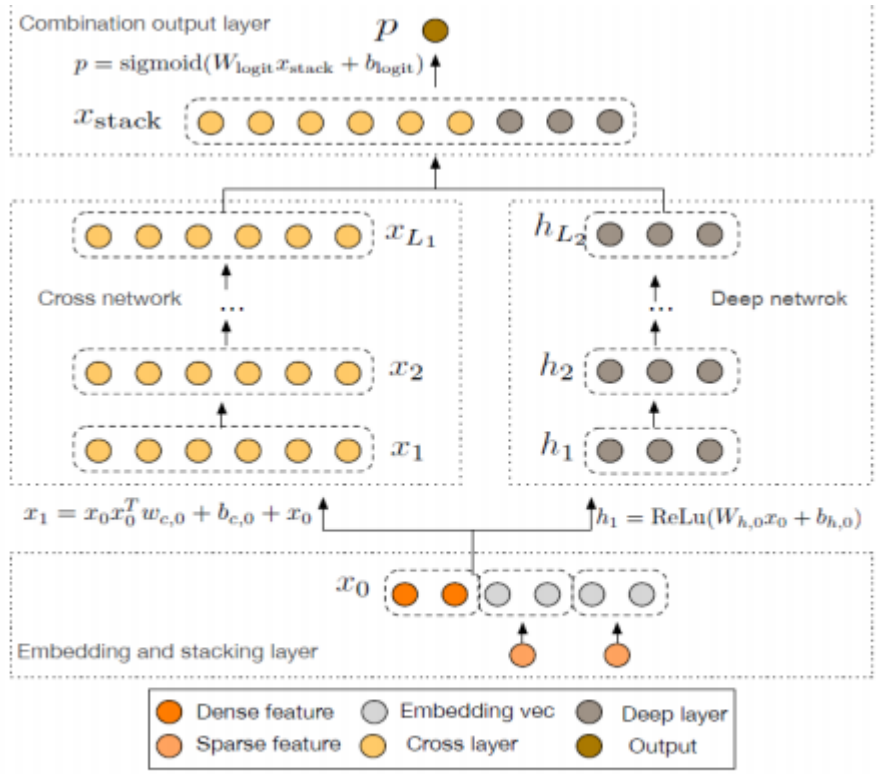
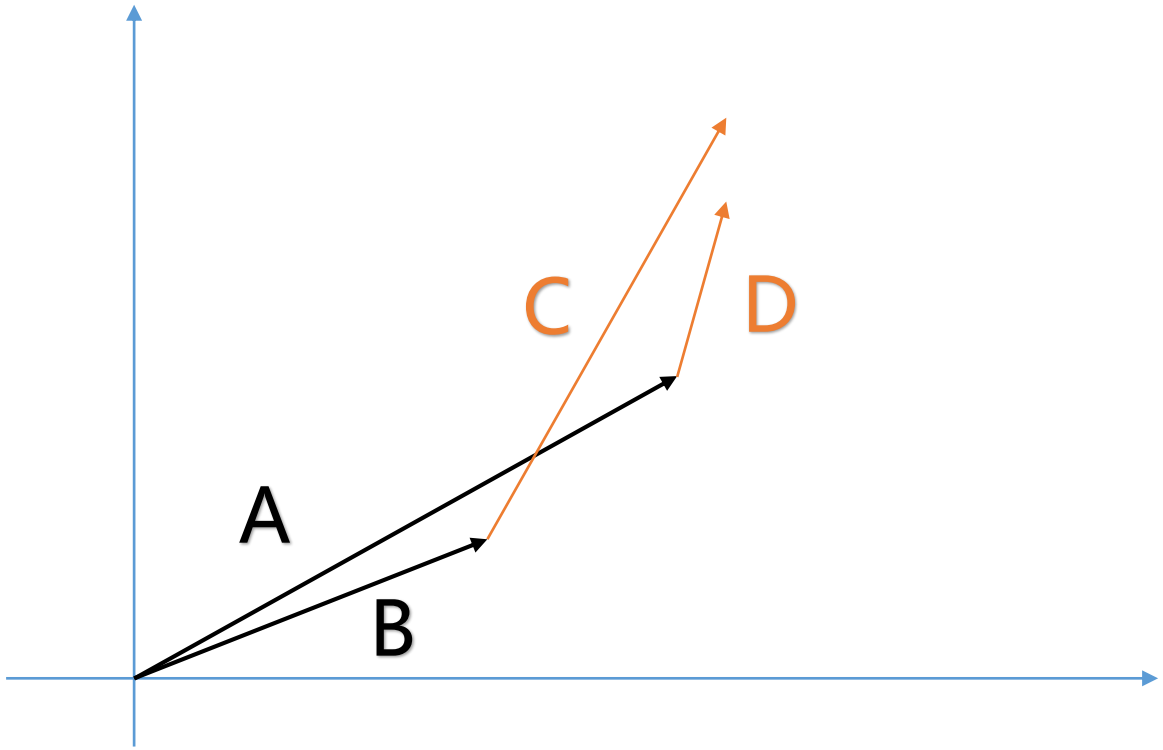


Figure 1: The Deep & Cross Network

参考论文：Deep & Cross Network for Ad Click Predictions，2017

辛普森悖论



已知：
A = a1/a2
B = b1/b2
C = c1/c2
D = d1/d2

现象：
A>B
D>C

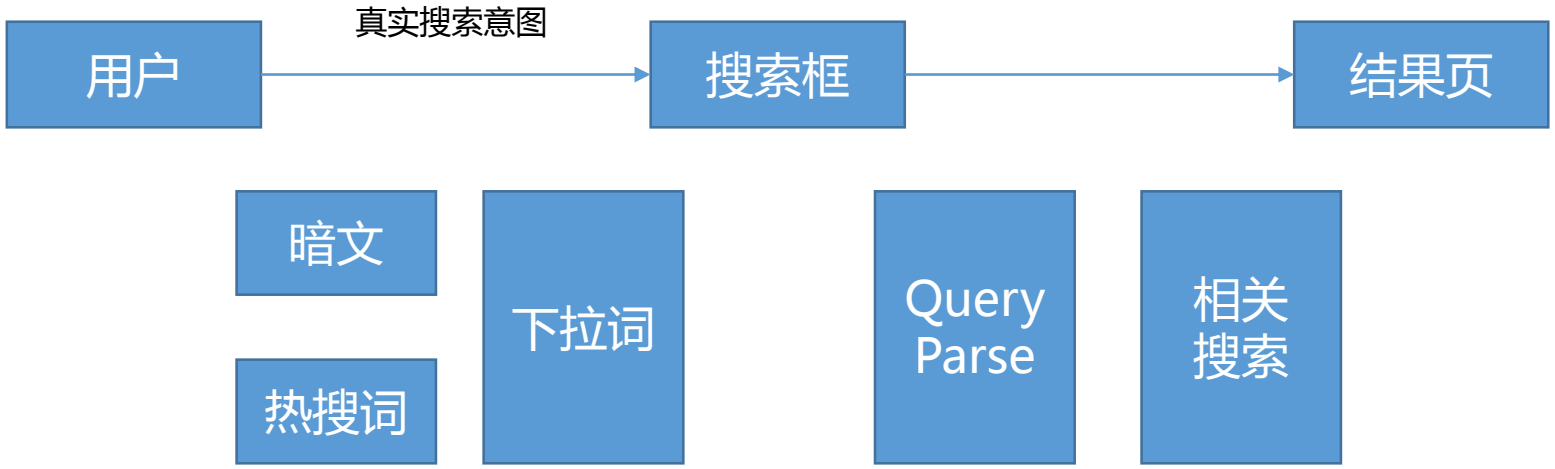
(a1+d1)/(a2+d2) <
(b1+c1)/(b2+c2)

JD搜索学习分享

xpguo

Reference: 《引流产品算法和框架》--谢群群

引流产品



业务

- 大促，满减，三免一
- 新品扶持
- 大促卖场入口

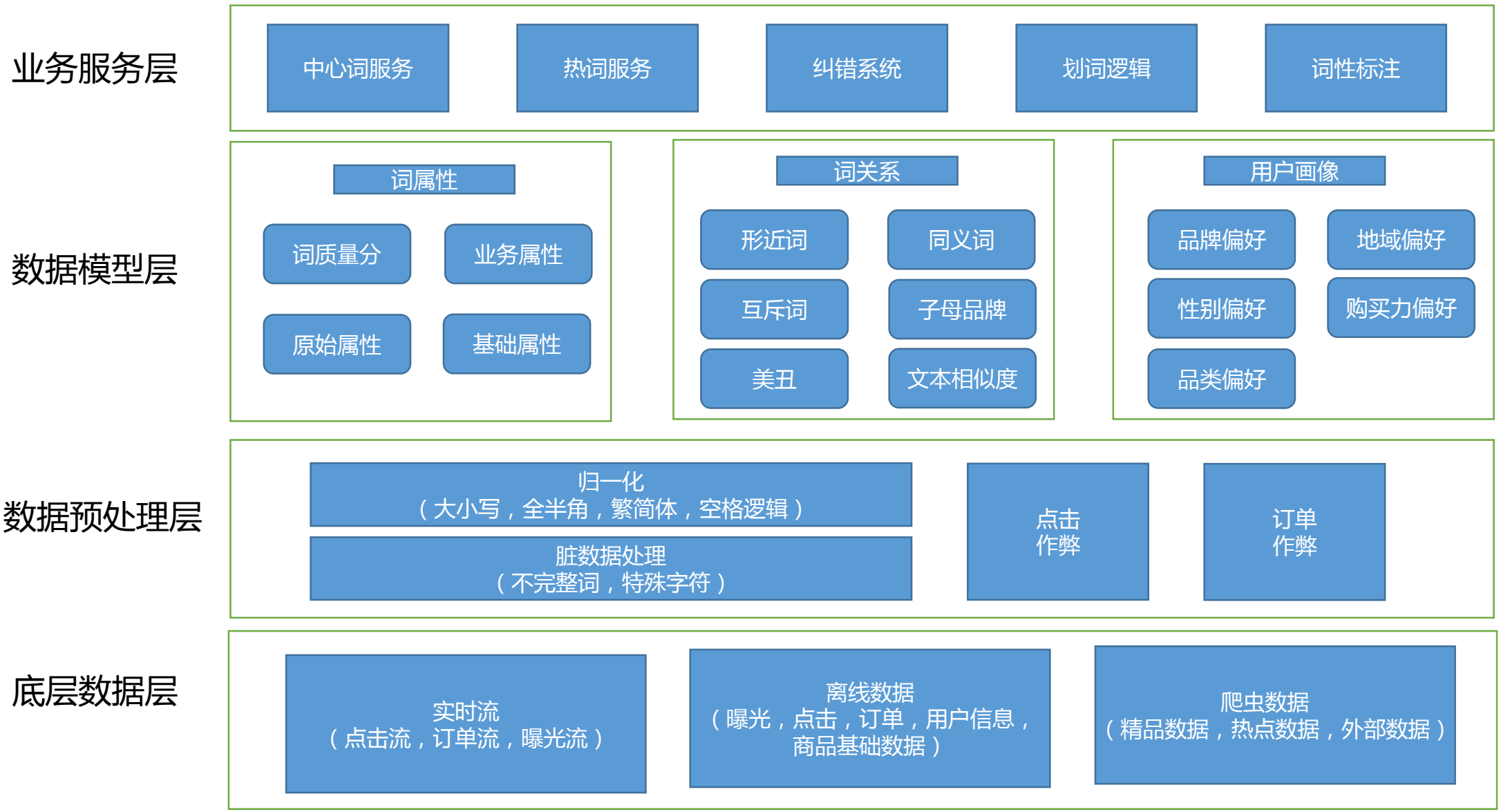
算法

- 引入KPI因子（转化率，客单价）
- 语义分析（送给爸爸的礼物）
- 序列识别（停止词，修饰词）
- 主题聚类（场景识别，育儿场景）
- 个性化模型（年龄，品牌，类目，历史词）

数据

- 归一化（大小写处理，简称）
- 命名实体识别
- 数据清洗（反作弊层）
- 违禁词（菜刀）
- 不完整词（阿迪达）
- 季节词提权（茶叶）
- 纠错

基础数据框架


业务服务层
中心词服务
热词服务
纠错系统
划词逻辑
词性标注
数据模型层
词属性
词质量分
业务属性
原始属性
基础属性
词关系
形近词
同义词
互斥词
子母品牌
美丑
文本相似度
用户画像
品牌偏好
地域偏好
性别偏好
购买力偏好
品类偏好
数据预处理层
归一化
（大小写，全半角，繁简体，空格逻辑）
脏数据处理
（不完整词，特殊字符）
点击
作弊
订单
作弊
底层数据层
实时流
（点击流，订单流，曝光流）
离线数据
（曝光，点击，订单，用户信息，
商品基础数据）
爬虫数据
（精品数据，热点数据，外部数据）

词画像系统

词性标注
- 产品词
- 品牌词
- 型号词/数量词
- 停用词

质量得分
- 规模指标
- 转化指标
- 综合指标

平台属性
- 大促词
- 违禁词

用户偏好
- 性别偏好
- 购买力
- 年龄

QueryParse

适合女生的热销新款手提电脑

词性标注
（适合|女生|的|热销|新款|手提电脑）
（停止词|修饰词|停止词|修饰词|修饰词|产品词）

个性化
（性别：女，0.97）
（购买力：high）
（品牌偏好：apple 0.6，sony 0.2）

类目预测
（电脑类目）

语义扩展
女生 or 超轻薄
手提电脑 or 笔记本

属性扩展
热销 -> 销量指标>X
新款 -> 上架时间> XXXX

场景扩展
场景：内容
主题：笔记本 女

QP框架

Understanding
- 词性标注
- 个性化
- 类目预测

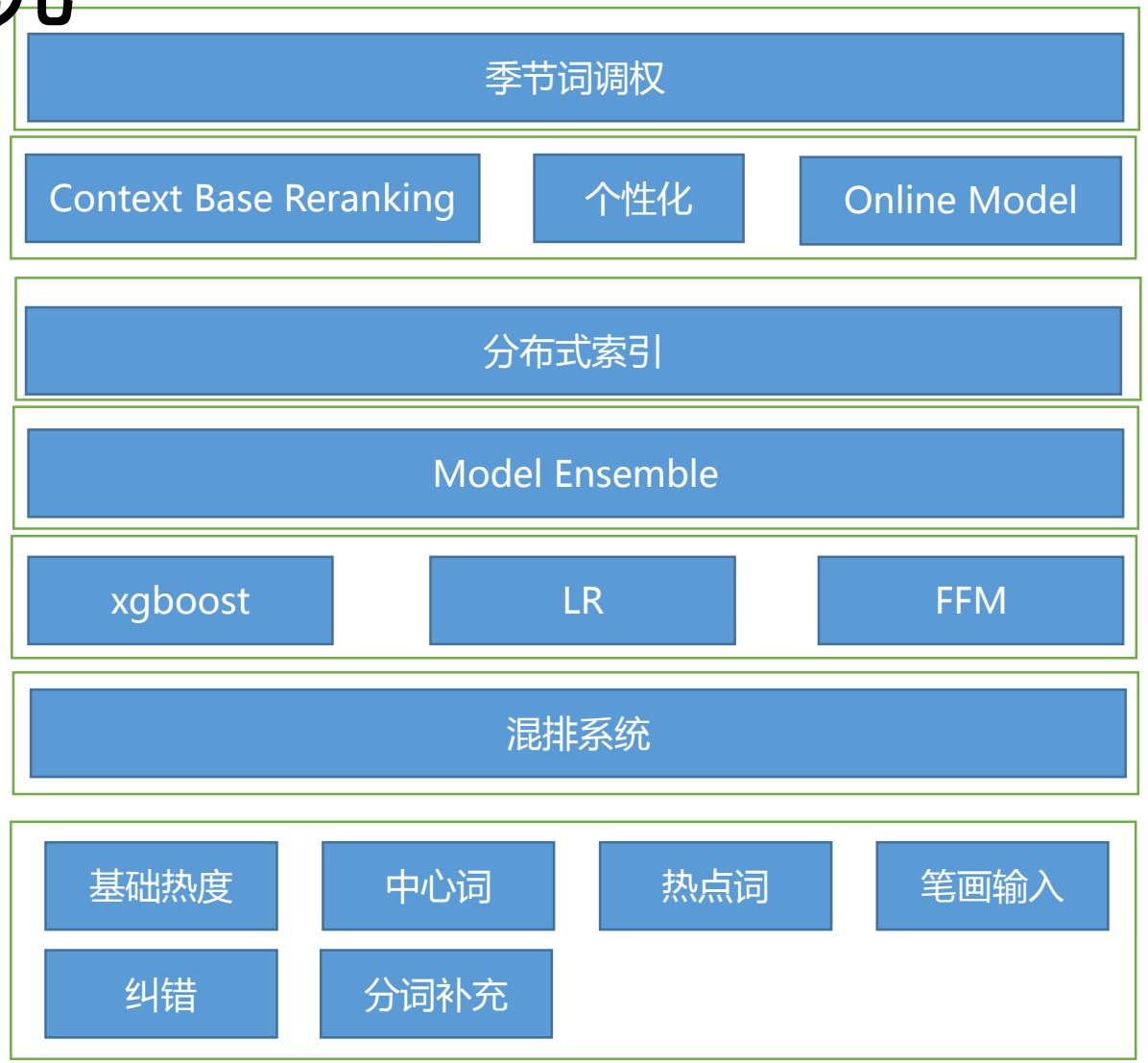
Rewrite
- 违禁词
- 大促词
- 业务改写
- 服务Box
- 纠错/划词

语法树

NLP
- 语义分析（命名实体识别）
- 序列识别
- 主题分析

实时计算
- Query标注
- Tag类目预测
- Query相似度

排序系统


季节词调权
Context Base Reranking
个性化
Online Model
分布式索引
Model Ensemble
xgboost
LR
FFM
混排系统
基础热度
中心词
热点词
笔画输入
纠错
分词补充