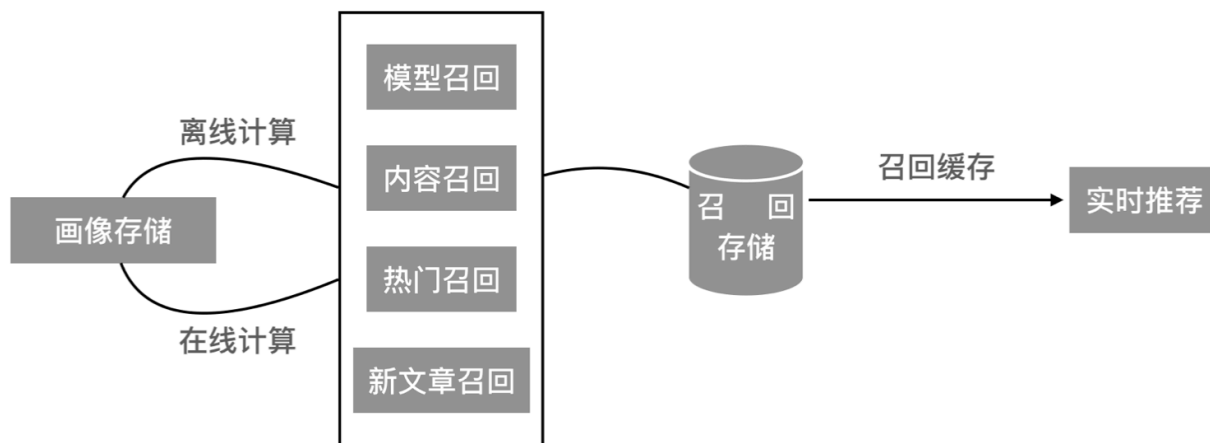




3.5 离线用户基于内容召回集

3.5.1 基于内容召回实现(基于画像召回)

- 目的：实现定时离线更新用户的内容召回集合
- 步骤：
 - 1、过滤用户点击的文章
 - 点击行为数据
 - 相似度计算的结果
 - 2、用户每次操作文章进行相似获取并进行推荐

3.6 离线用户召回定时更新

3.7 离线排序模型训练

3.6.1 离线排序模型-CTR预估

CTR (Click-Through Rate) 预估：给定一个Item，预测该Item会被点击的概率

3.6.2 排序模型

最基础的模型目前都是基于LR的点击率预估策略，目前在工业使用模型做预估的有这么几种类型

- 宽模型 + 特征工程
 - LR/MLR + 非ID类特征(人工离散/GBDT/FM)
 - spark 中可以直接使用
- 宽模型 + 深模型
 - wide&deep, DeepFM
 - 使用TensorFlow进行训练

- 深模型：
 - DNN + 特征embedding
 - 使用TensorFlow进行训练

3.6.3 特征处理原则

3.6.5 spark LR 进行预估

- 2、读取用户点击行为表，与用户画像和文章画像，构造训练样本
 - 预测某个用户对某篇文章的点击率预估
 - 特征值+目标值，拿到样本?? 行为数据
- 1、需要通过spark读取HIVE外部表，需要新的sparksession配置
 - 增加HBASE配置
- 2、读取用户点击行为表，与用户画像和文章画像，构造训练样本
 - 注意：toDF(['user_id', 'birthday', 'gender', 'article_partial'])
 - ValueError: Some of types cannot be determined by the first 100 rows, please try again with sampling
 - 缺失值太多。无法自动决定类型
 - 20个文章关键词权重
- 3、LR模型进行训练，LR模型预测、结果评估

离线计算分析：

- 用户画像：保存到hbase user_profile, HIVE 关联user_profile_hbase
- spark sql: user_profile_hbase

```
("hbase.zookeeper.quorum", "192.168.19.137"),
("hbase.zookeeper.property.clientPort", "22181")
```

3.6.5 点击率预测结果

- `probability`：对于每个样本，，和不点击（0为目标）的概率
 - 0.9, 0.1
 - 文章点击(1为目标)的概率

• 3.6.6 模型评估-Accuracy与AUC

• AUC：模型样本不均衡现象

- 0.5 ~ 1
- from pyspark.mllib.evaluation import BinaryClassificationMetrics
- AUC: 0.7364334522585716
 - from pyspark.mllib.evaluation import BinaryClassificationMetrics

- 输入的类型：RDD
- 取出res_transfrom的结果当中，模型预测点击的概率
def get_click_p(row): return float(row.clicked), float(row.probability[1])
- row.probability[1]): 预测点击的概率，都比较小

```
accuracy_score(arr[:, 0], arr[:, 1].round())
0.9051438053097345
```

```
roc_auc_score(arr[:, 0], arr[:, 1])
0.719274521004087
```

- LR模型
- 推荐： 1 [123, 345], 两个样本，预测

3.8 离线ctr特征中心更新

- 了解特征服务中心的作用
- 用户特征、文章的特征更新到HBASE中
 - 为实时推荐中召回结果排序快速提供特征，构造样本

```
create 'ctr_feature_user', 'channel'

create 'ctr_feature_article', 'article'
```

3.8.2 用户特征中心更新

- 目的：计算用户特征更新到HBASE

```
create 'ctr_feature_user', 'channel'
4                                     column=channel:13, timestamp=1555647172980,
value=[ ]
```

3.8.2 文章特征中心更新

article_id	channel_id	weights	articlevector
26	17	[0.19827163395829...	[0.02069368539384...
[17.0,0.198271633...			

存入： article_id, features

3.8.3 离线特征中心定时更新

离线部分

```
# 添加一个文章画像更新任务，每个一个小时运行一次更新
scheduler.add_job(update_article_profile, trigger='interval', hours=1)
# 添加一个定时的用户画像的更新任务，每隔2个小时更新一次用户画像
scheduler.add_job(update_user_profile, trigger='interval', hours=2)
# 添加一个定时运行用户召回结果的任务，每隔3小时新推荐一次用户的召回池
scheduler.add_job(update_user_recal, trigger='interval', hours=3)
# 添加一个定时运行特征中心的更新任务，每隔4小时新推荐一次用户的召回池
scheduler.add_job(update_feature, trigger='interval', hours=4)
```

4.1 实时计算业务介绍

实时流处理概述

- 实时计算：数据的即时处理
- 流式计算：数据是持续不断的输入的

4.1.1 实时计算业务需求

- 解决用户冷启动问题
- 实时计算能够根据用户的点击实时反馈，快速跟踪用户的喜好
- flume + kafka + spark streaming：搭建实时系统

4.1.2 黑马推荐业务图