Hadoop在大型内容推荐系统中的应用

蔡迎东

QQ: 1170625333 Weibo : 蔡王仔

分享内容

- 背景
- 推荐效果
- 技术选型
- 技术实现
- 推荐系统的评测
- Hadoop&Hive使用经验
- 下一步工作

背景

■ 需求

- ✓ 网易门户新闻数量急剧膨胀，但是新闻利用率很低:

  网易门户每天新发布的文章数量约为10万篇， 但是有PV的文章不足10%。大量的文章成为长尾而沉没，得不到展示的机会。

- ✓ 用户期望在web端和移动端能即时快捷地看到自己感兴趣的文章和话题。

背景

■ 面临的挑战

- ✓ 用户访问量大，每天产生的原始日志文件大小约为500G，日志数量约为10亿多条。

- ✓ 文章数量大，系统每天新增约10万篇文章，2万个左右的图片和短视频。

- ✓ 时效性要求高，热点新闻发布后需要迅速在推荐区域推荐出来以及Push到移动端。

分享内容

- 背景
- 推荐效果
- 技术选型
- 技术实现
- 推荐系统的评测
- Hadoop&Hive使用经验
- 下一步工作

推荐效果实例1-奥运个性化新闻推荐

■ 奥运期间的体育个性化新闻推荐

London 2012 诸强争霸 更多>>

- [关注] 奥运八强成难逾越天堑 波尔出局中国男单获利好
- [人物] 日本偶像被封最萌男主播 樱井翔多栖发展很全面
- [动态] 日本申诉“求”来一块银牌 内村向乌克兰说抱歉
- [人物] 孙杨将父母介绍给朴泰桓认识 敌意全无互相欣赏
- [人物] 皇帝甜瓜入"3届奥运"俱乐部 这一纪录恐再无来者

系列1

- [声音] 法国总统：不想再失败 巴黎或申办2024年奥运会
- [花絮] 美国杂志评奥运性感明星 孙杨引“花痴”追捧
- [花絮] 英自行车一姐身材火辣 科比：我最想看她比赛
- [动态] 西班牙国奥队零进球出局 马卡报:足球没有感情
- [人物] 华裔小将邢延华险爆冷赢李晓霞 获世界首富夸奖

系列2

编辑推荐

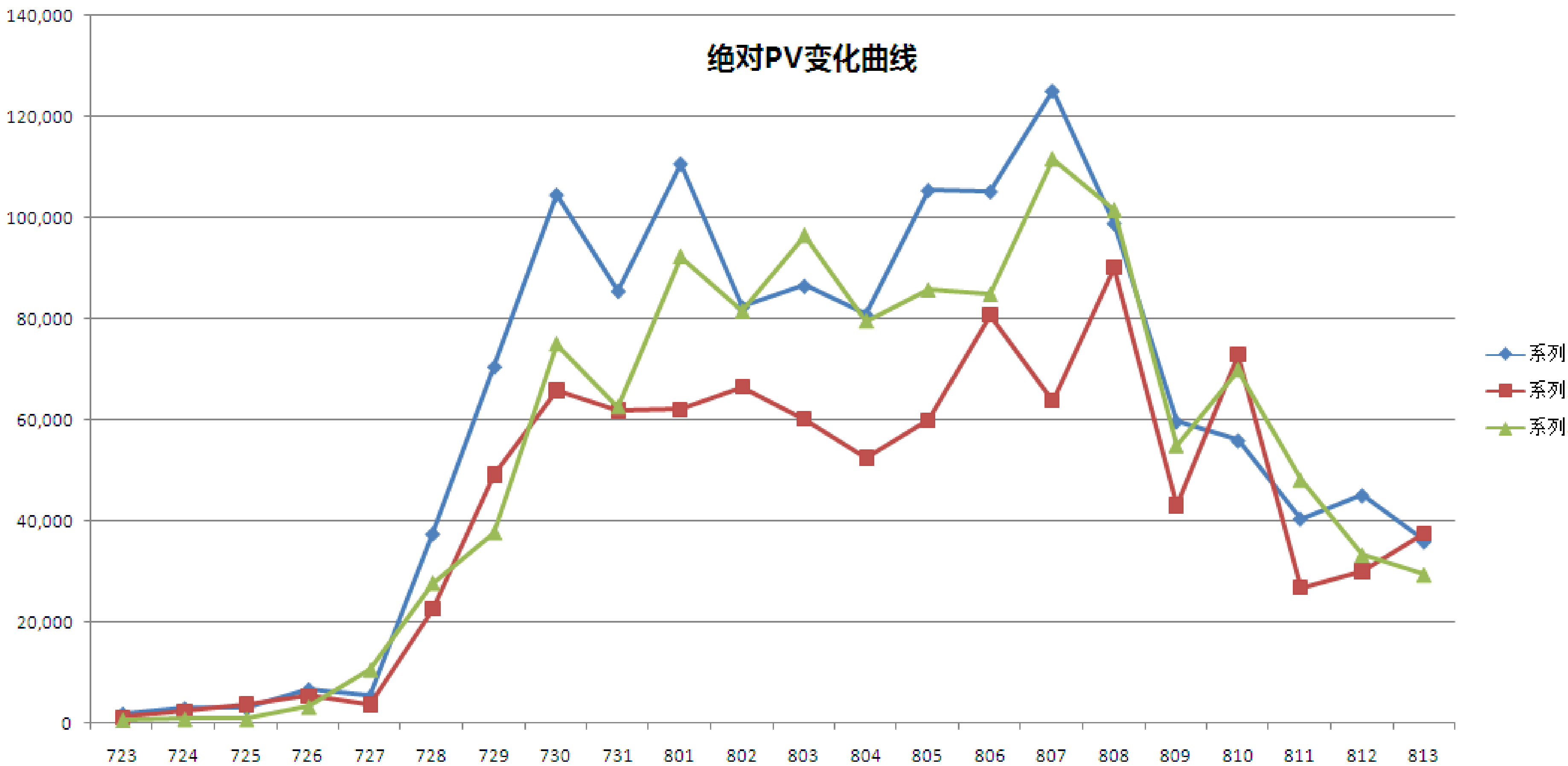
- [推荐] 纳瓦罗旧伤复发将休战1场 西班牙冲金前景蒙阴影
- [推荐] 小AI伦敦遇"劫匪"被吓跑 直言梦十不在最佳状态
- [推荐] 反击质疑 中国代表团:怎么没人怀疑菲鱼用兴奋剂
- [推荐] 曾经外媒为中国夺金不安 而今中国形象已经转变
- [推荐] 瑞士球员微博发歧视言论被除名:我犯下一个大错

系列3

推荐系统

推荐效果实例1-奥运个性化新闻推荐

- 与邻近的编辑推荐区的绝对PV对比图



推荐效果实例1-奥运个性化新闻推荐

■ 与邻近的编辑推荐区的日均PV对比

位置	上线前	上线后	相对PV（上线前/上线后）
	日均PV	日均PV	
系列1	1313	78,246	59.6
系列2	1050	55,593	52.9
系列3	718	65,818	91.7

综合来看，处于页面板块下方的个性化推荐区域的PV接近或者超过了上方的两块编辑推荐区，推荐区域日均PV的增长速度约为编辑推荐区的1.63倍。

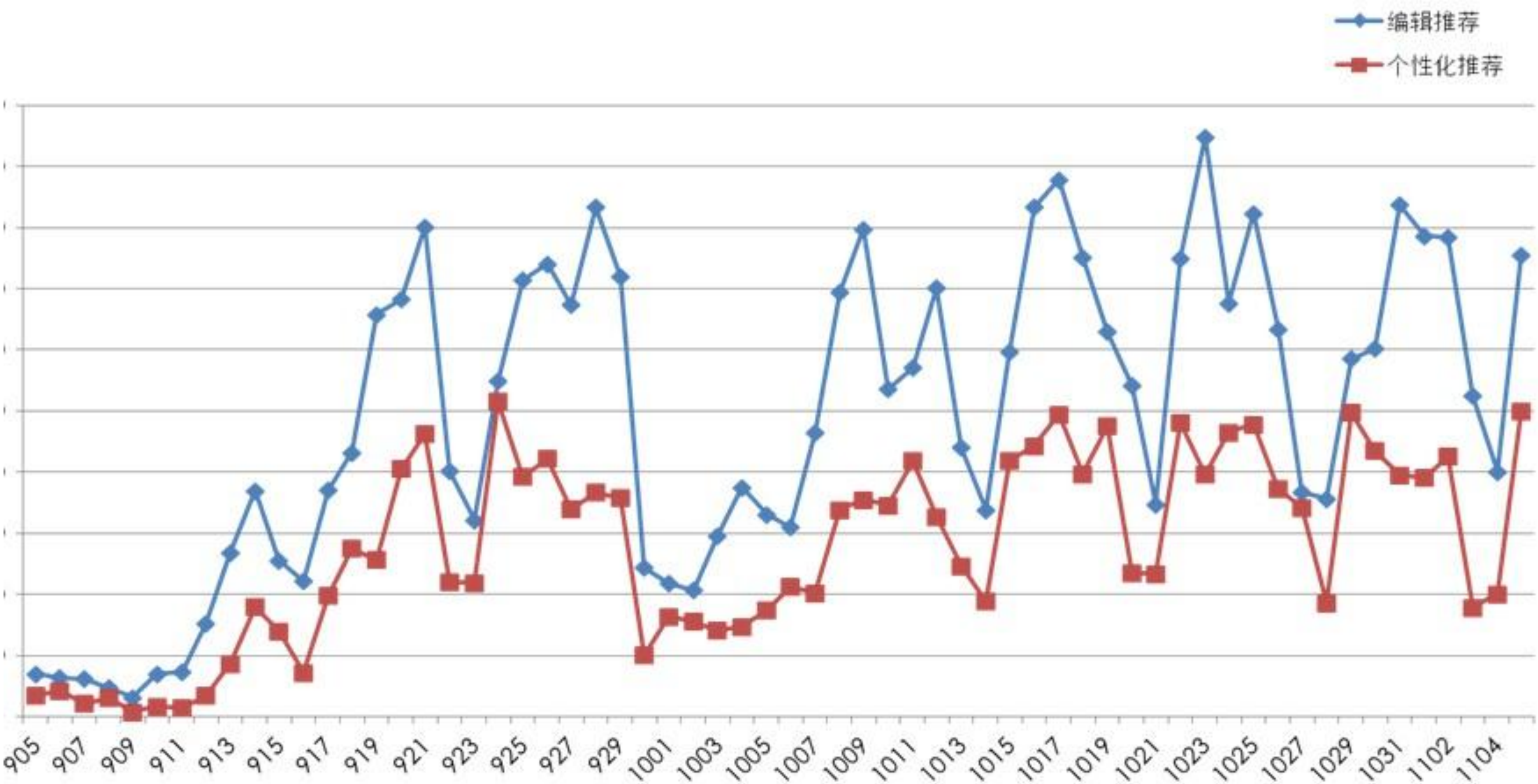
推荐效果实例2-邮箱首页个性化新闻推荐

- 网易邮箱首页的个性化新闻推荐

编辑推荐区 **系统推荐区**

推荐效果实例2-邮箱首页个性化新闻推荐

- 与左侧的编辑推荐区的绝对PV对比



推荐效果实例2-邮箱首页个性化新闻推荐

■ 与左侧的编辑推荐区的日均PV对比

位置	相对PV (上线后/上线前)
编辑推荐区	9.31
个性化推荐区	13.15

- ✓ 上线后个性化推荐区域的PV增长较快，为左侧的编辑推荐区域的增长速度的1.41倍。
- ✓ 上线前个性化推荐区的排行榜约占全部邮箱新闻PV的27.4%，上线后个性化推荐区占总PV比提高至35.6%。

分享内容

- 背景
- 推荐效果
- 技术选型
- 技术实现
- 推荐系统的评测
- Hadoop&Hive使用经验
- 下一步工作

技术选型

- 基于内容的推荐系统
  - ✓ 对用户(User)和物品(Item)分别建模。
  - ✓ 计算物品和用户的模型的相似度。
  - ✓ 把和用户的模型相似度最高的物品推荐给用户。
- 基于协同过滤的推荐系统
  - ✓ 与系统的业务无关。
  - ✓ 不是根据用户和物品本身的属性，而是根据用户的访问记录来挖掘出相似度。

技术选型

- 协同过滤的优点
  - ✓业务无关
  - ✓ 算法实现和基础数据采集相对简单
  - ✓ 业界广泛采用，比如电商网站

- 但是……

技术选型

- 考虑到新闻自身的特点，放弃协同过滤
  - ✓ 协同过滤是基于访问记录进行推荐，只有被人访问过的文章才能被推荐出来，这对时效性要求比较高的新闻推荐是严重的缺陷。
  - ✓ 新闻的生命周期很短，会造成访问记录的极度稀疏，这给根据访问记录来计算相似性带来了很大的困难。

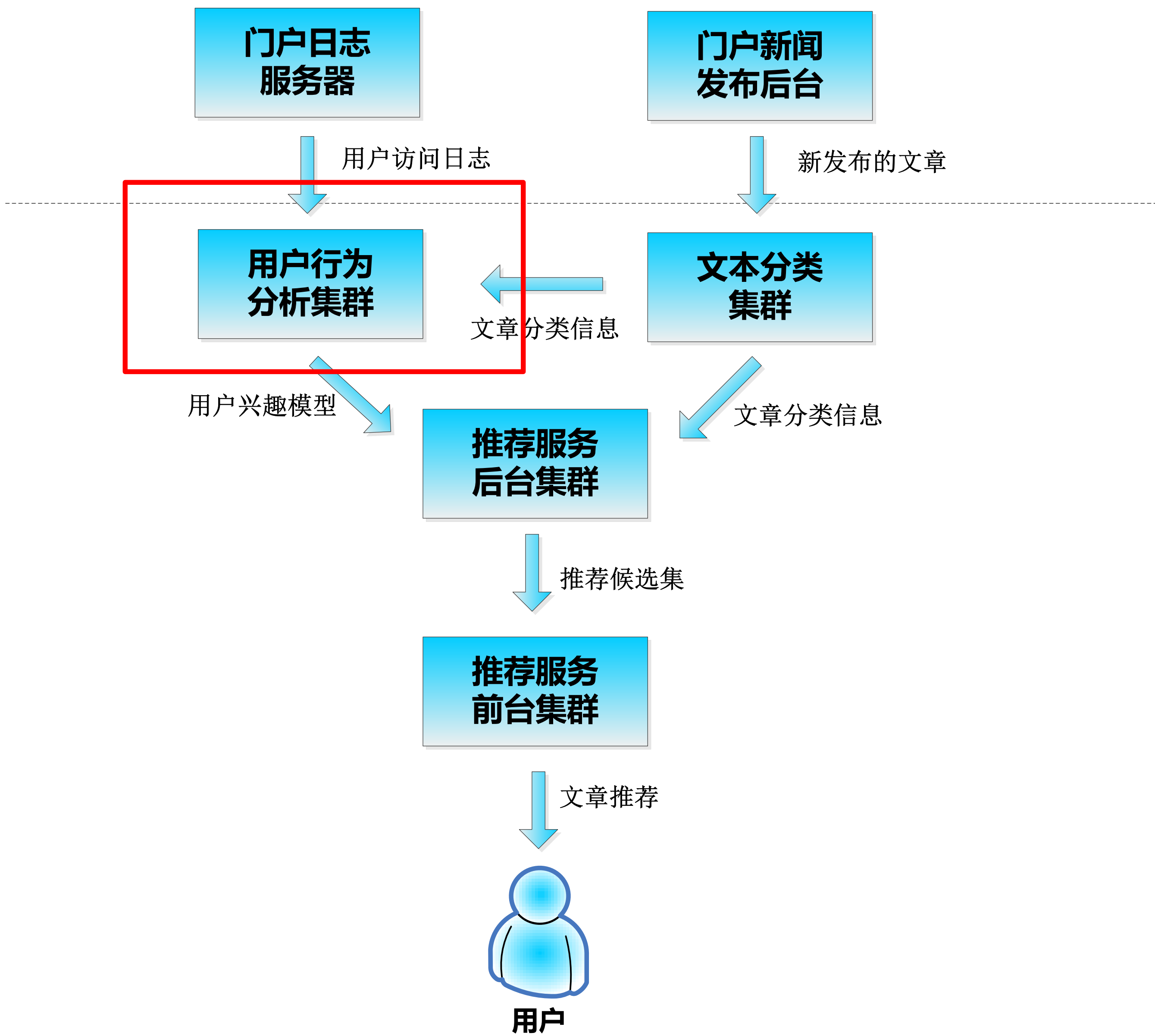
技术选型的结果

- 新闻推荐
  - ✓ 基于内容的推荐
- 图集和视频推荐
  - ✓ 基于协同过滤的推荐

分享内容

- 背景
- 推荐效果
- 技术选型
- 技术实现
- 推荐系统的评测
- Hadoop&Hive使用经验
- 下一步工作

系统总体架构



门户日志
服务器
门户新闻
发布后台
用户访问日志
新发布的文章
用户行为
分析集群
文本分类
集群
文章分类信息
用户兴趣模型
文章分类信息
推荐服务
后台集群
推荐候选集
推荐服务
前台集群
文章推荐
用户

用户行为分析集群

- 从门户的用户访问日志中挖掘出用户的兴趣，构建用户的兴趣模型。

- 采用Hadoop&Hive作为数据挖掘工具。

描述用户兴趣的粒度

- Tag（标签）
  - ✓ 门户全站的文章涉及的范围很广，用关键词或标签来描述用户的兴趣显得太细。
- Category（类别）
  - ✓ 对全站的频道进行细分，每个频道进一步划分成多个类别。用户兴趣的粒度以类别为主。
- Topic（话题）
  - ✓ 对文章数量较大的类别，采用介于标签和类别之间的话题对类别进行进一步的细分。

用户兴趣模型的表示-用户兴趣向量

- 用户兴趣向量
  - ✓ 用户兴趣向量的维度是文章的类别。
  - ✓ 每一维的值表示用户对这一类文章的兴趣比重。

	国内	财经	汽车
用户兴趣向量 →	(20% ,	20% ,	60%)
	对国内文章的兴趣比重占20%	对财经文章的兴趣比重占20%	对汽车文章的兴趣比重占60%

用户兴趣向量计算的演进

- 最初的计算方法：
  以用户的点击分布作为用户的兴趣向量

用$interest(u,c_i)$表示用户u对类别为c_i的文章的兴趣，

$D(u,c_i)$表示用户u的点击分布中维度为c_i的值，

$$interest(u,c_i) = D(u,c_i) = \frac{N(u,c_i)}{N(u)},$$

其中：

$N(u,c_i)$ = 用户u点击过的类别为c_i的文章数

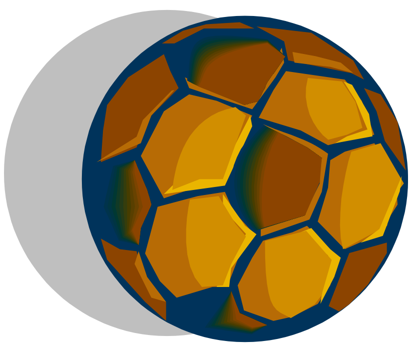
$N(u)$ = 用户u点击过的文章总数

用户兴趣向量计算的演进

- 以点击分布作为兴趣向量的问题
  - ✓ 在奥运期间推荐区域效果显著，但是在奥运结束后一段时间内推荐区域的PV下降明显。
- 原因分析
  - ✓ 对系统推荐的文章进行统计后发现，在奥运结束后，向大多数用户推荐的文章仍然以体育类的文章为主。

影响用户点击行为的因素

- 用户浏览了一篇足球类别的文章……
  - ✓ 可能他本身是个足球迷。
  - ✓ 也可能他是个伪球迷，只是受了当前热点事件的影响，比如世界杯、欧洲杯和奥运会。



影响用户点击行为的因素

用户的点击行为受用户自身的真实兴趣和新闻热点两个因素的影响，点击分布本身无法准确地反映用户的真实兴趣。

热门 VS. 冷门

- 对于不同的文章类别，用户的点击行为应该具有不同的权重。
  - ✓ 用户阅读了一篇国内新闻类文章…
  - ✓ 用户阅读了一篇亲子类文章…
  - ✓ 国内新闻类的文章很热门，多数人每天都会看看
  - ✓ 亲子类文章比较冷门，只有特定的人群才会关注
  - ✓ 亲子类文章比国内新闻类的文章更能反映用户的兴趣

用户对冷门类别的文章的点击行为应该具有更高的权重。

用户兴趣向量计算的演进

- 改进办法：应用贝叶斯公式描述用户兴趣向量，平衡点击分布中新闻热点因素的影响

用户u对类别为c_i的文章的兴趣$interest(u,c_i)$可以描述为$p(click \mid category=c_i)$，根据贝叶斯公式:

$$interest(u,c_i)=p(click \mid category=c_i)$$

$$=\frac{p(category=c_i \mid click)\,p_t(click)}{p_t(category=c_i)} \propto \frac{D(u,c_i)}{D(c_i)}$$

其中:

$p(category=c_i \mid click)$近似为用户$u$的点击分布$D(u,c_i)$

$p(category=c_i)$近似为整体用户的点击分布$D(c_i)$

用户兴趣向量计算的演进

■ 改进后的公式的效果

- ✓ 用整体的点击分布**D(c_i)**平衡单个用户的点击分布**D(u,c_i)**中新闻热点这个因素的影响，更能反映用户的真实兴趣。
- ✓ 用整体的点击分布**D(c_i)**对在门户中占主导地位的国内新闻等大文章分类的权重进行了惩罚，使得推荐的结果更具有多样性。

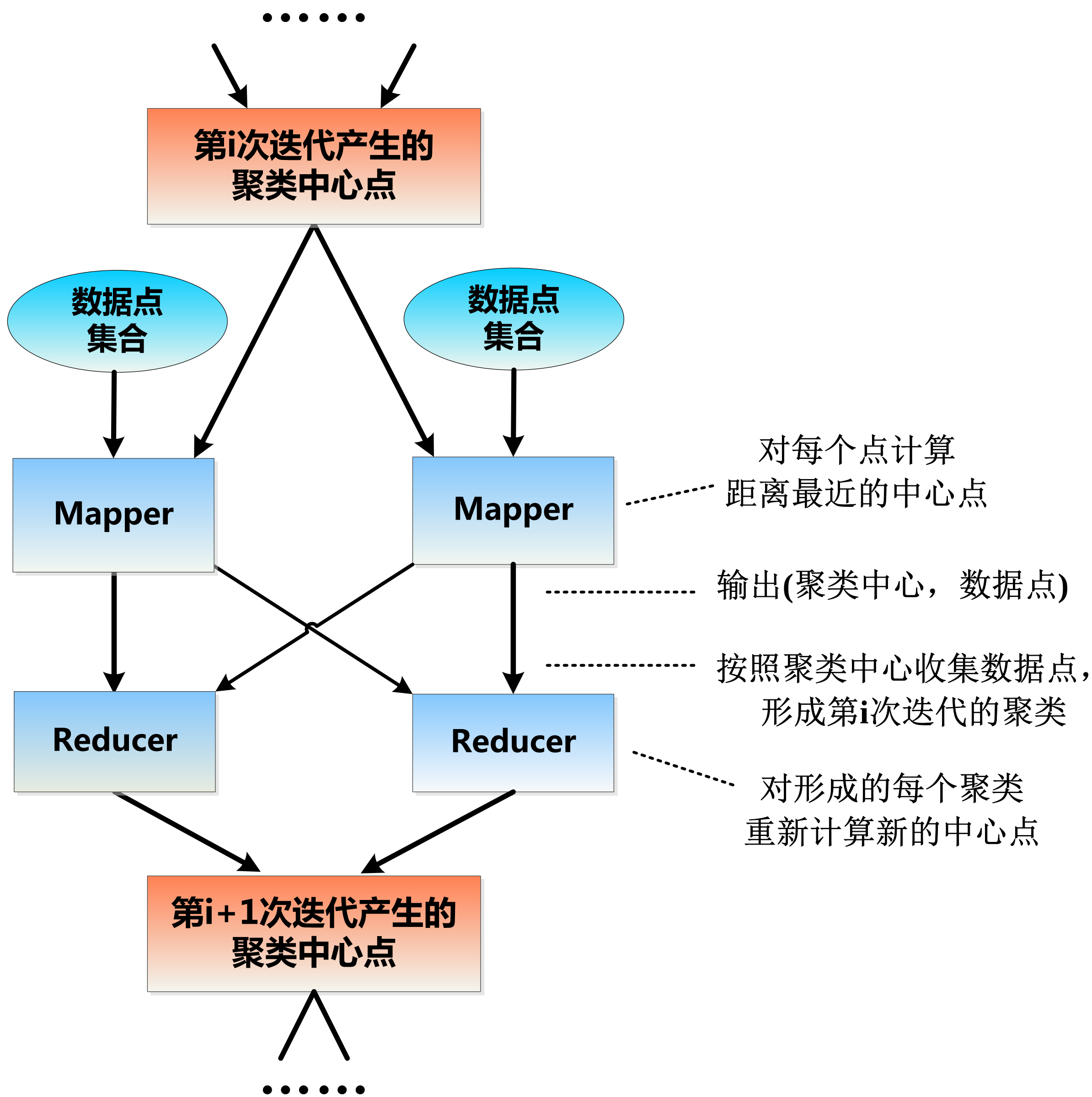
$$interest(u,c_i)=\frac{D(u,c_i)}{D(c_i)}$$

改进的算法上线之后，推荐区域的PV逐渐回升。

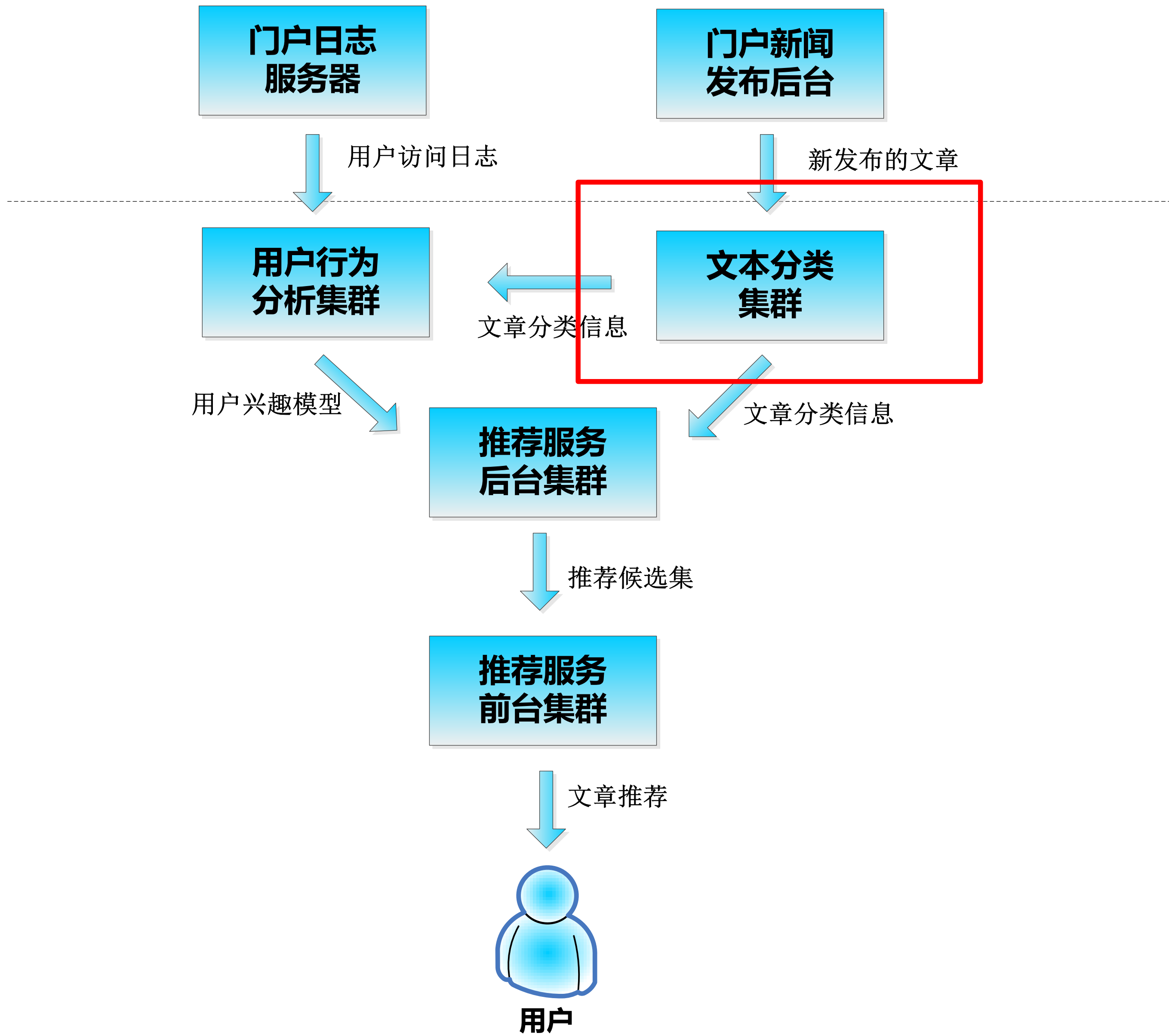
用户兴趣向量的聚类

- **聚类**
  - ✓ 将兴趣向量相似的用户聚成一个用户类，同一个聚类的用户看到的推荐文章是相同的。
- **作用**
  - ✓ 降维，避免对每个用户都进行推荐运算。
  - ✓ 把海量的用户降维成有限的几个聚类之后，也使得预计算成为可能。
- **算法**
  - ✓ 采用基于**MapReduce**实现的**K-Means**算法实现聚类。

基于MapReduce的K-Means算法



系统总体架构


门户日志
服务器
门户新闻
发布后台
用户访问日志
新发布的文章
用户行为
分析集群
文本分类
集群
文章分类信息
用户兴趣模型
文章分类信息
推荐服务
后台集群
推荐候选集
推荐服务
前台集群
文章推荐
用户

文本分类集群

- 实现对门户新闻的自动分类。
- 为用户兴趣分析提供基础数据，系统以文章的类别为粒度来描述用户的兴趣。
- 为推荐服务集群的推荐预计算提供分类文章的候选集。

文本分类的预处理

- 中文分词
  - ✓ 对门户编辑历年积累的文章关键词进行整理，丰富中文词库。
  - ✓ 定期用热门文章的关键词对中文词库进行更新。
- 特征词提取
  - ✓ TF*IDF模型
  - ✓ 词的位置信息
  - ✓ 词的信息增益

文本分类算法

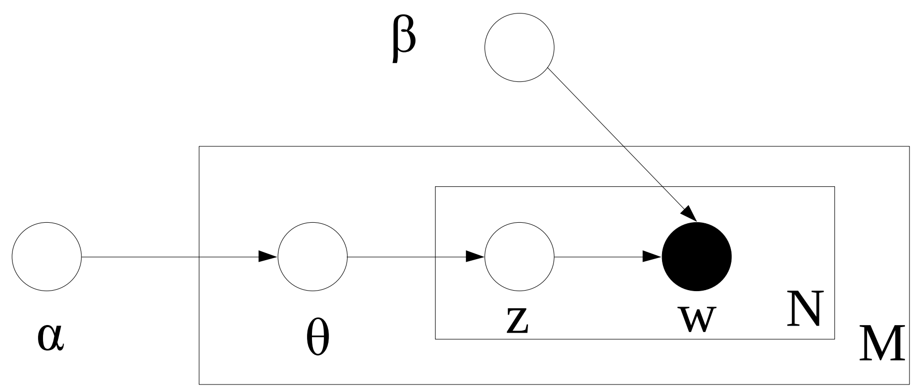
- 常用文本分类算法
  - ✓ k-NN
  - ✓ 朴素贝叶斯
  - ✓ 支持向量机
  - ✓ 最大熵
- 选择多项式模型的朴素贝叶斯算法
  - ✓ 实现简单，分类和训练的速度都很快。
  - ✓ 尽管特征独立性假设比较粗糙，但是对分类来说已经足够了。

分类算法的问题

- 监督模型的局限性
  - ✓ 需要大量的人工标注语料。
  - ✓ 类别需要预先指定。
  - ✓ 类别数不能过多，类别数越多，分类的效果越差。
- 解决办法
  - ✓ 对文章数量多、内容丰富的大分类，采用无监督的话题模型作为补充。

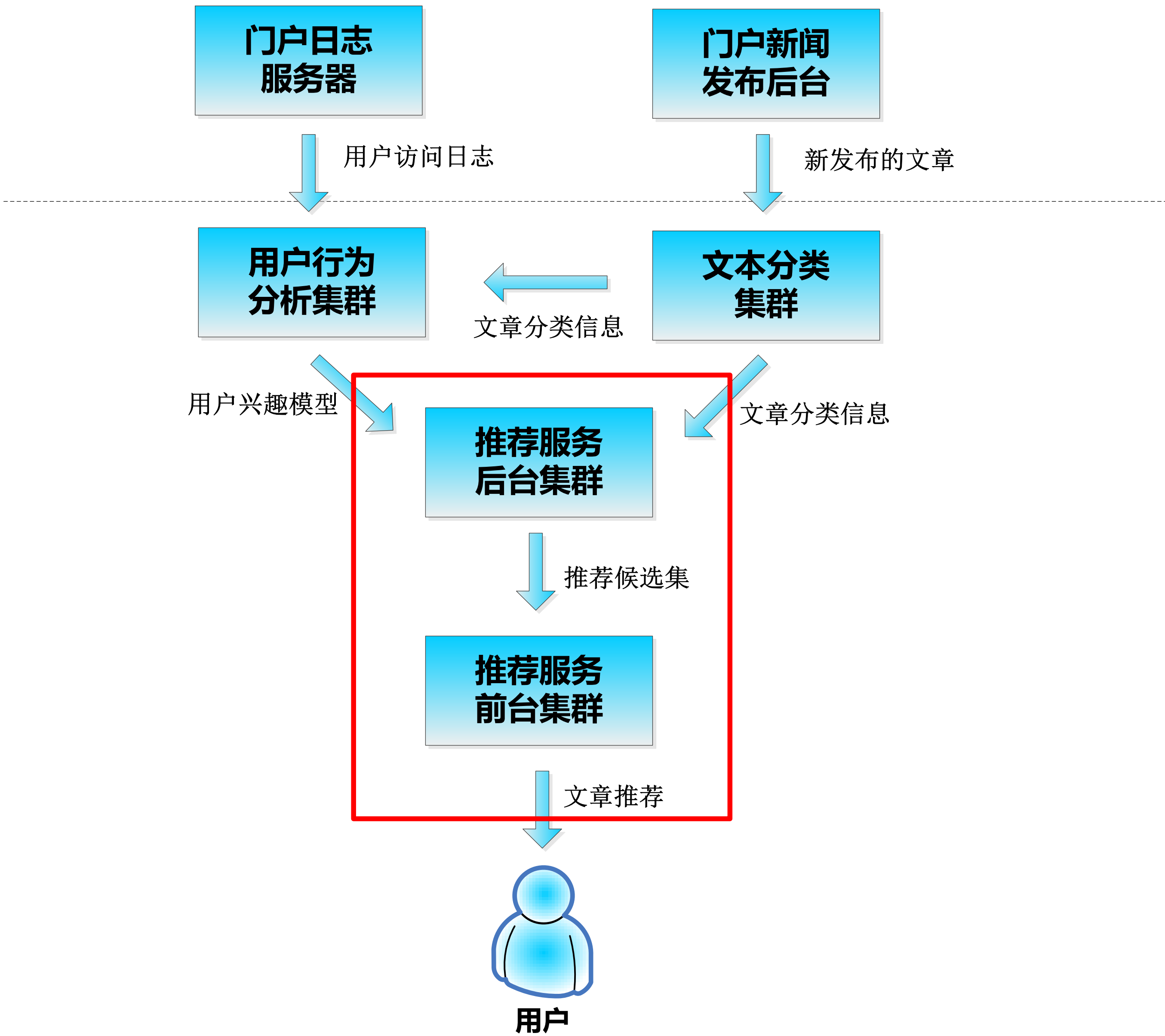
话题模型

- 常用话题模型
  - ✓ LSA
  - ✓ PLSA
  - ✓ LDA



- 选择LDA
  - ✓ 生成式模型
  - ✓ 话题：词的分布概率
  - ✓ 文档集：由这些内在的话题生成
  - ✓ 文档生成：对于每一个词的位置，先选定一个话题，然后在话题里面选择词去填充

系统总体架构



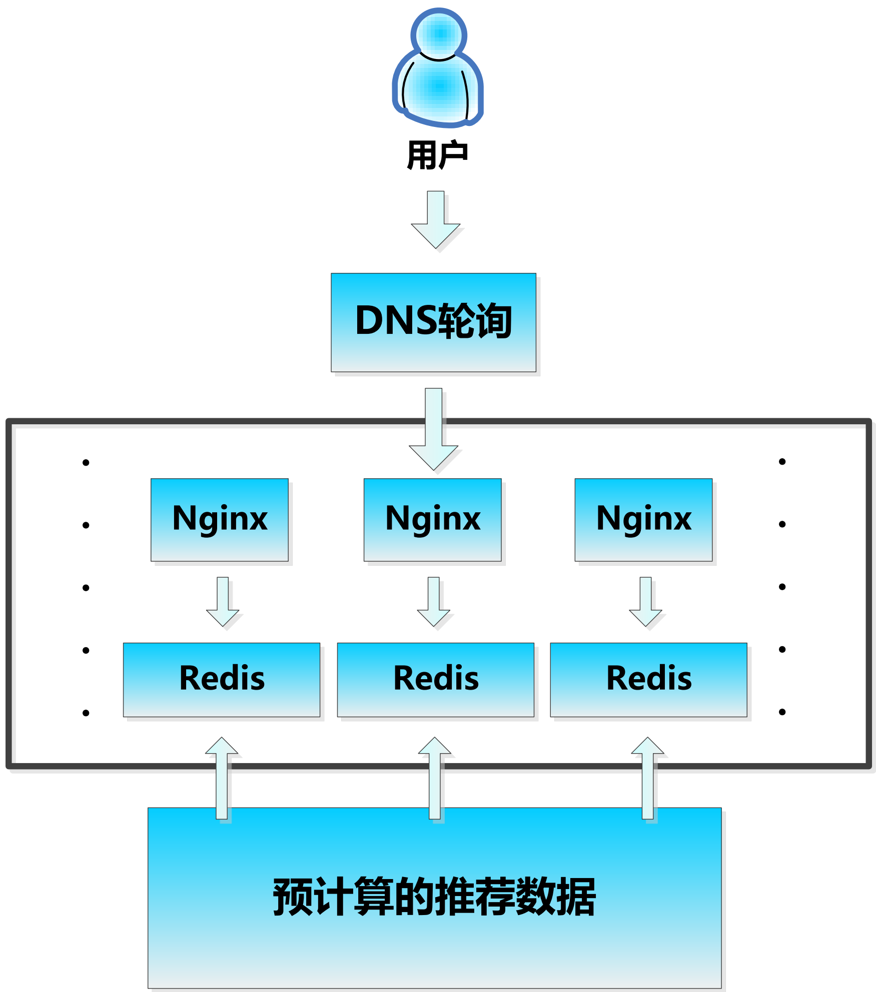
门户日志服务器
门户新闻发布后台
用户访问日志
新发布的文章
用户行为分析集群
文本分类集群
文章分类信息
用户兴趣模型
文章分类信息
推荐服务后台集群
推荐候选集
推荐服务前台集群
文章推荐
用户

新闻推荐服务后端集群

- 推荐的预计算
  - ✓ 融合用户行为分析和文本分类的运算结果。
  - ✓ 为每个文章类别生成候选集。
  - ✓ 对每个用户兴趣向量的聚类集合，从候选集中按照兴趣向量中的类别比例配置文章，组合成推荐结果。

新闻推荐服务前端集群

- 对外提供推荐服务: Nginx+Redis



协同过滤

- 对图集和视频推荐采用协同过滤
  - ✓ 图集和视频的数量远小于用户数量，选择Item-based协同过滤。
  - ✓ 度量Item和Item的相似：访问过Item A和Item B的User的重合度越大，则认为Item A和Item B越相似。
- 采用基于MapReduce模型的Item-based协同过滤算法。

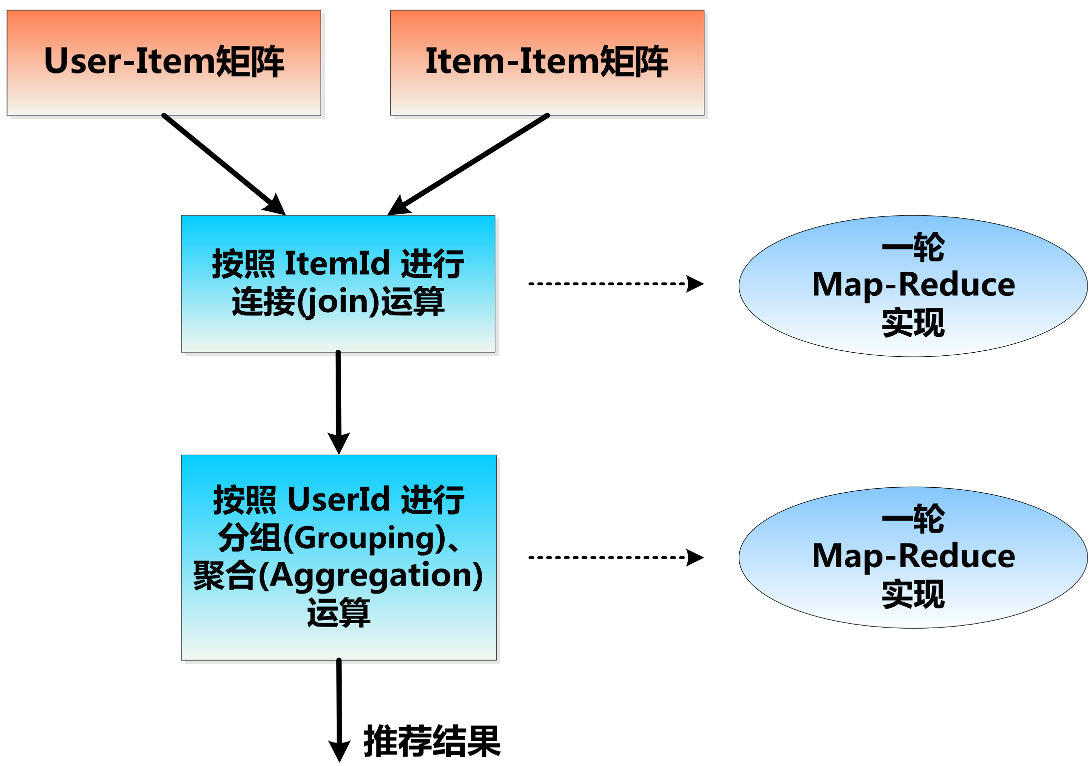
基于MapReduce的Item-based协同过滤

■ 矩阵运算

- ✓ Item-base算法用数学语言来描述就是：

  User-Item矩阵 × Item-Item相似矩阵 。

- ✓ 稀疏矩阵的乘法是MapReduce的经典应用之一。



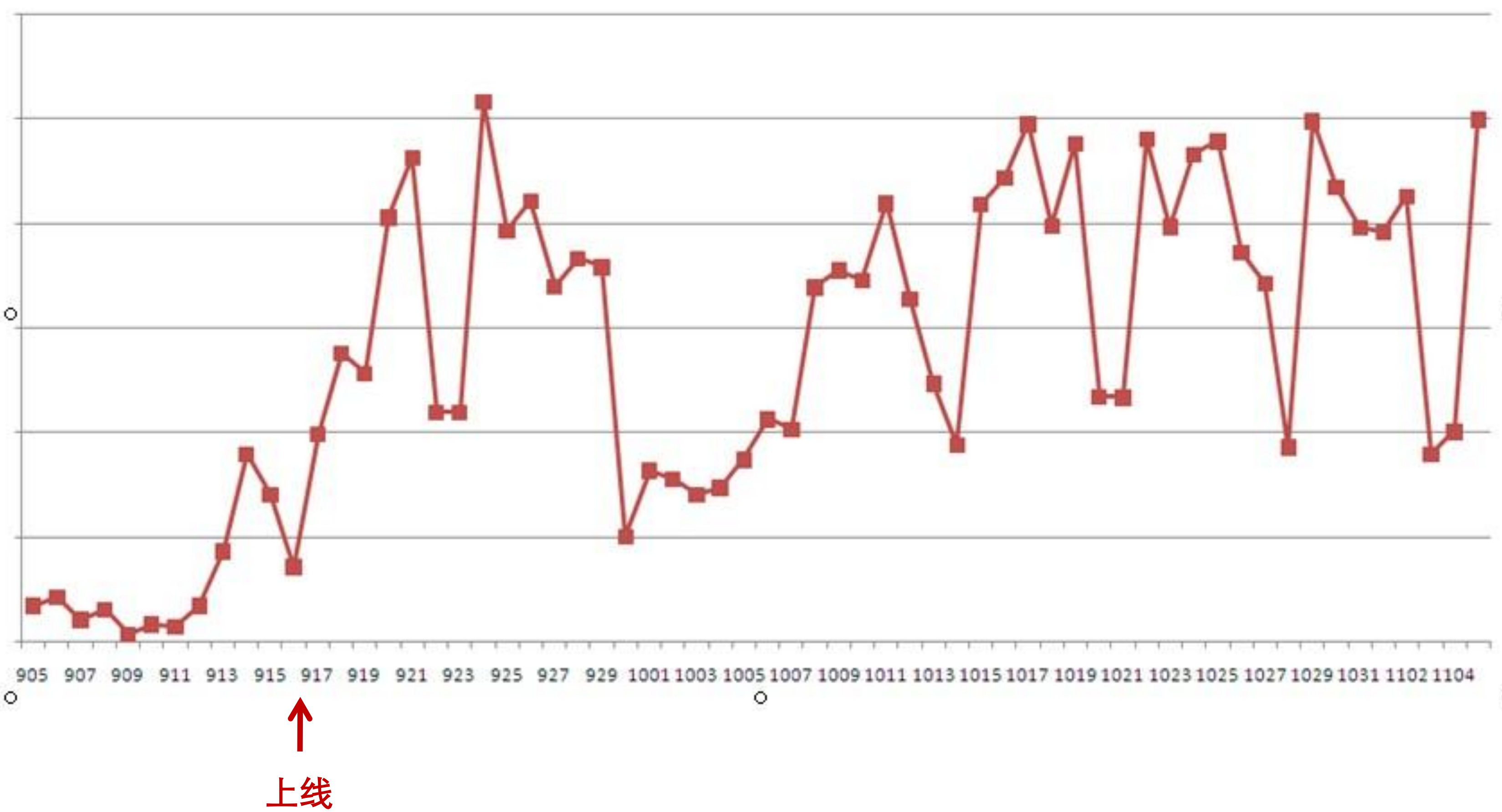
分享内容

- 背景
- 推荐效果
- 技术选型
- 技术实现
- 推荐系统的评测
- Hadoop&Hive使用经验
- 下一步工作

推荐系统的在线评测方法

- 同区域对比上线前后的绝对PV

邮箱首页个性化新闻推荐的绝对PV对比图



推荐系统的在线评测方法

- 与邻近区域对比上线前后的相对PV

奥运个性化新闻推荐的相对PV对比

位置	上线前	上线后	相对PV（上线前/上线后）
	日均PV	日均PV	
系列1	1313	78,246	59.6
系列2	1050	55,593	52.9
系列3	718	65,818	91.7

推荐系统的在线评测方法

■ ABTest: 在线对不同的算法进行效果对比

✓ 通过一定的规则将用户随机分成几组。

✓ 对不同组的用户采用不同的算法。

✓ 通过对不同组的用户的各种评测指标进行对比，来评价不同的算法的效果。

项目	日均PV	展示比例	折算后PV
以用户的点击分布作为兴趣向量的算法	155,546	33%	466,638
改进之后的算法	432,237	66%	648,355

推荐系统的离线评测方法

- 离线评测
  - ✓ 准确率
  - ✓ 召回率
  - ✓ 覆盖率
  - ✓ 多样性
- 离线评测的结果可能与线上的指标之间存在差异，算法的最终效果还是得参照线上指标。

分享内容

- 背景
- 推荐效果
- 技术选型
- 技术实现
- 推荐系统的评测
- Hadoop&Hive使用经验
- 下一步工作

Hadoop & Hive 使用经验

- 瓶颈
  - ✓ 磁盘I/O
  - ✓ 网络I/O
- 解决办法
  - ✓ 使用LZO或者SNAPPY进行压缩。
  - ✓ 使用combiner等方法减少每个mapper的输出。
  - ✓ 使用MultiOutputFormat输出多种结果，避免多次计算。

Hadoop & Hive 经验

- 进行数据统计时，使用Hive可以显著减少工作量，提高工作效率。
- 对同一个表进行多种统计操作时，使用multi-insert能大大提高运算速度。
- 使用JDBC方式驱动Hive时，如果Hive所在机器负载比较大时，容易导致网络IO错误，导致运算结果丢失，应减少Hive服务器的任务量，降低负载。
- 使用ganglia和nagois做好监控，对错误及时报警。

分享内容

- 背景
- 推荐效果
- 技术选型
- 技术实现
- 推荐系统的评测
- Hadoop&Hive使用经验
- 下一步工作

下一步工作

- 进一步完善用户模型
  - ✓ 人口学信息
  - ✓ 时间因素、地域因素
- 对文章分类的进一步细分
  - ✓ 不同的频道采用不同的分类粒度
- 融合多种推荐策略
  - ✓ 内容过滤+协同过滤
- 完善推荐系统的评价体系