BDTC
2014 中国大数据技术大会
BIG DATA TECHNOLOGY CONFERENCE
暨第二届CCF大数据学术会议

基于全网内容的新闻客户端推荐系统

刘佳

搜狐搜狐移动新媒体中心
移动研发部主管

liujiaplus@163.com
weixin: liujiaatict

信息量大，屏幕小，我该怎么办？

RSS 订阅？ 太高端

搜索引擎？ 不知道搜啥

。 。 。

推荐引擎！说曹操曹操到

让适合的资讯在适合的时间以适合的方式传递给适合的用户

内容推荐引擎

上搜狐，知天下

MRD的主要工作

推荐引擎、APP换量、定向投放

推荐的计算量：

2亿用户 X 100万资讯，每日700GB实时用户行为日志，8秒入库时延。每秒6万条入库数据，近200万次入库操作。每秒10000次推荐请求,95%<50ms,0. 5%>1s

用户粘性：

天人均210篇文章消费

推荐转化率：15%

时效性：

用户反馈时效性保证

资讯的时效性保证

移动端新闻推荐特点

新闻推荐系统

广告系统

搜索引擎

前者是精准广告

终极追求：转化率、辅助指标ROI、用户效果

内容推荐：终极追求：良好的用户体验（模糊）

推荐人均消费次数、人均阅读时长、人均消费数量、人均分享评论、推荐转化率。

后者是搜索引擎

区别：搜索引擎包括对内容理解、内容爬取、文本关键词主题提取、文本分类、主题分类、内容索引、垃圾过滤、page rank、反作弊。用户输入消歧、关联检索。

内容推荐：

用户分群、用户画像、兴趣反馈校正、内容反馈校正。

推荐类型比率反馈修正（兴趣发散程度不同、用户疲劳程度匹配）

缺少内容关联信息（PR）、内容冷启动情况再严重。

由于有用户输入，搜索Rank更重要。

由于丰富的用户行为，分类更重要。

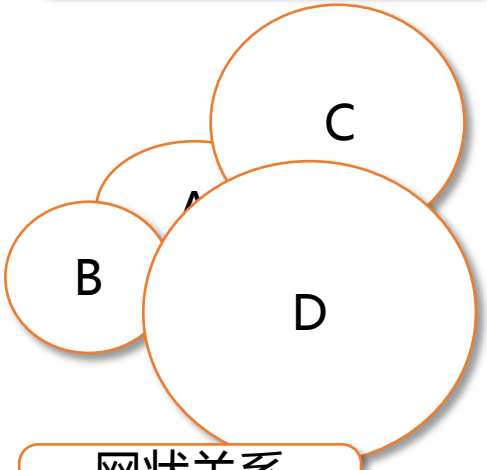
目前内容来源

- 自媒体约17000家 10到15万/每天
- 机构媒体+搜狐集团　　10到15万/每天
- 短视频 300万
- 搜狗内容50万/每天 去重后约 7万/每天
- 搜狗微信公众号 2万/每天

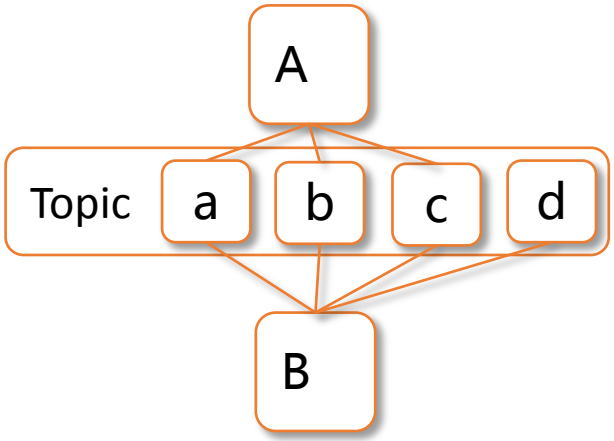
内容分类体系


树形关系
频道 13
生活
财经
财经
时政
。。。
子频道 56
旅游
美食
星座
摄影
。。。
Topic 5000
爸爸去哪
智能硬件
央行降息
。。。
标签词 6万
田亮
搜狐
微信
。。。
关键词 90万
娱乐
机会
投资
。。。

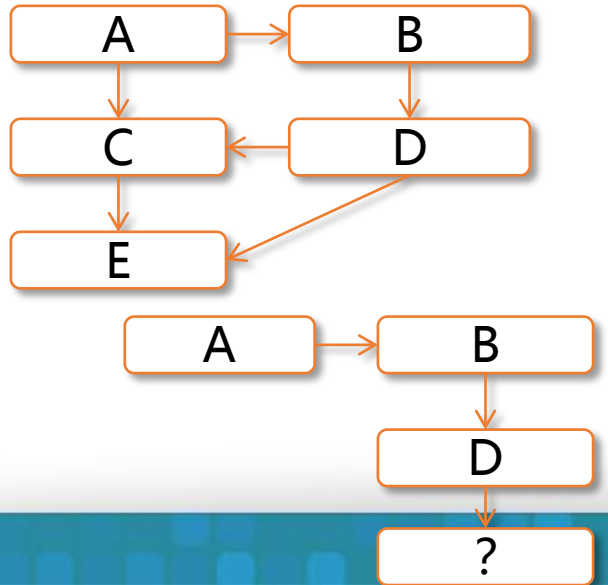
网状关系 Item Base Collaborative Filtering



网状关系 User Base Collaborative Filtering



网状关系 Context Tree



其他垂直规则

来源

地域：本地新闻

内容类型：视频、音频、游戏

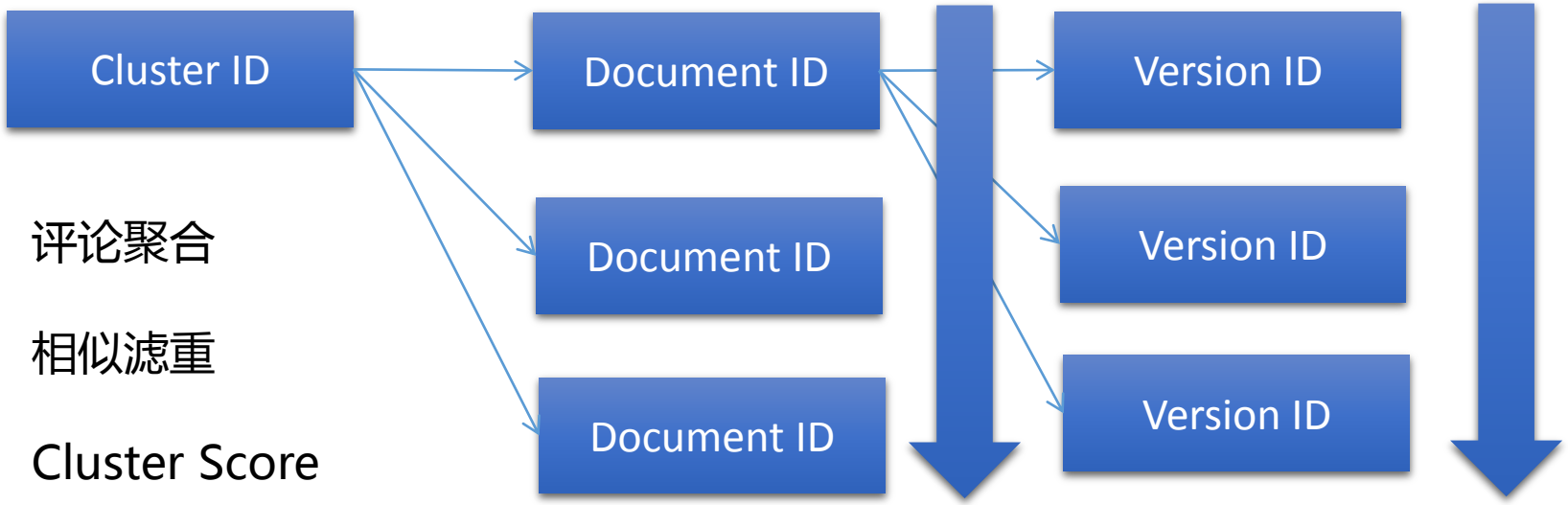
敏感程度：三俗、政治、社会

运营干预逻辑

新闻入库过程

1：内容同步、抽取	2：基于正文内容征生成全局ID	3：基于标题、摘要、关键词生成cluster ID
每日100万资讯内容 过滤垃圾信HTML标签、广告、页面重复内容保留 ---------- 面包屑 发布时间 来源 标题 正文 摘要 图片 评论	基于正文内容过滤重复，海明哈希 同步到各CMS生成全局ID 40万内容	基于标题及摘要关键词生成cluster Id 决定cluster score 根据版权、合作关系、来源质量、发布时间选择代表文章 20万Cluster ID

新闻入库过程



评论聚合

相似滤重

Cluster Score

静态指标：

图片数

来源PV

时间

动态指标：

评论数

PV

转化率

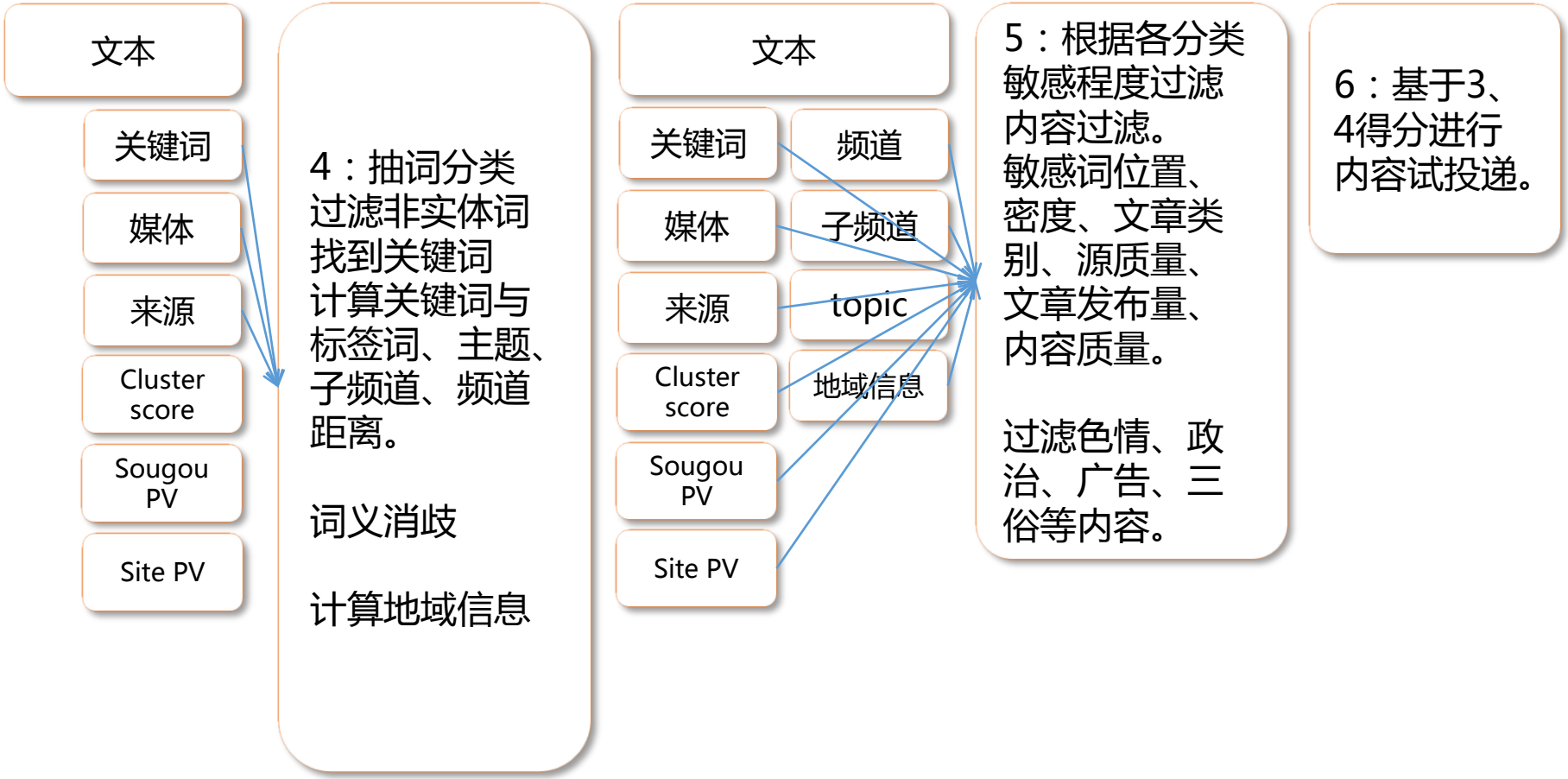
静态指标：

版权

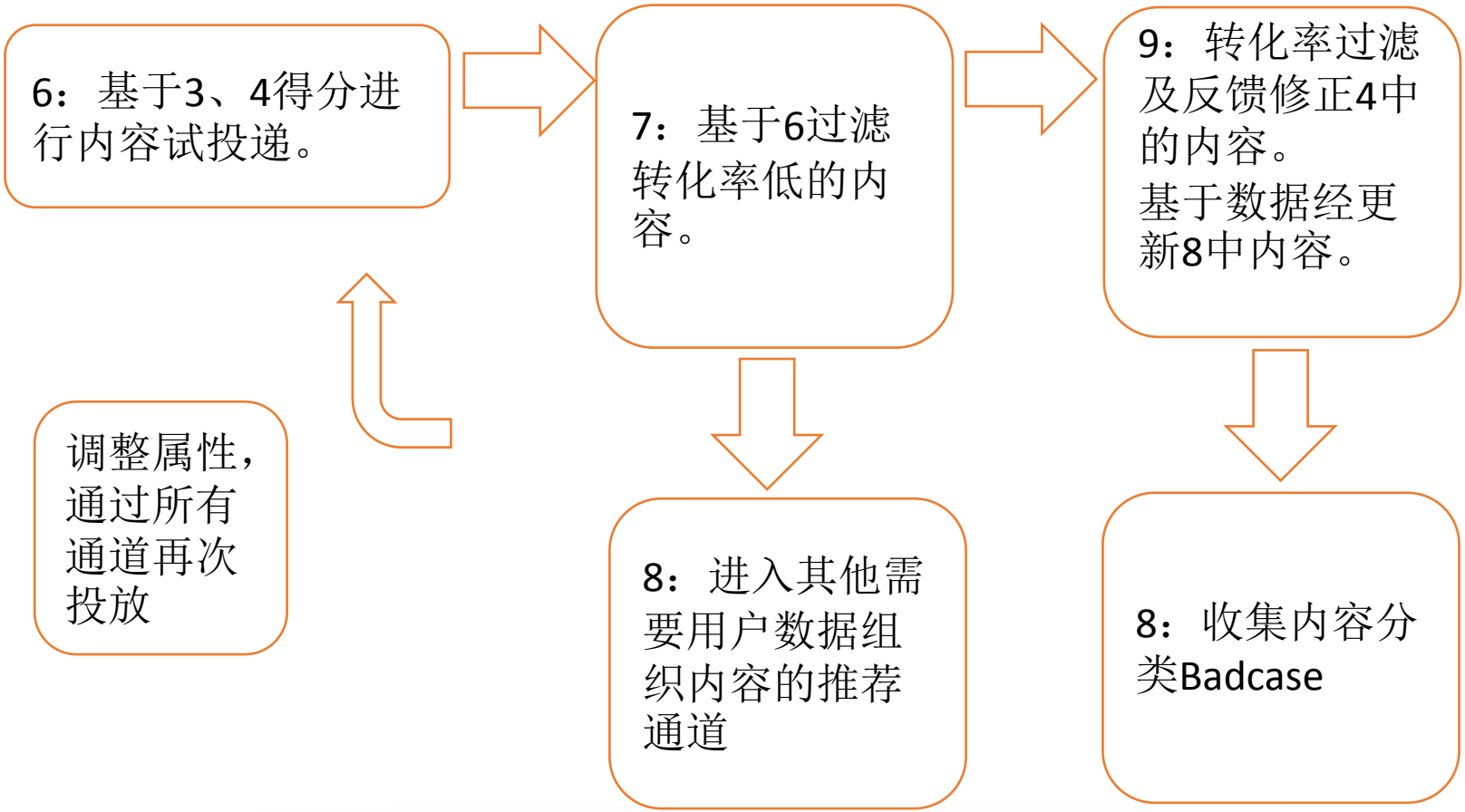
来源PV

时间

新闻分类过程


文本
关键词
媒体
来源
Cluster score
Sougou PV
Site PV
4：抽词分类
过滤非实体词
找到关键词
计算关键词与标签词、主题、子频道、频道距离。
词义消歧
计算地域信息
文本
关键词
媒体
来源
Cluster score
Sougou PV
Site PV
频道
子频道
topic
地域信息
5：根据各分类敏感程度过滤内容过滤。
敏感词位置、密度、文章类别、源质量、文章发布量、内容质量。
过滤色情、政治、广告、三俗等内容。
6：基于3、4得分进行内容试投递。

新闻投递过程


6：基于3、4得分进行内容试投递。
7：基于6过滤转化率低的内容。
9：转化率过滤及反馈修正4中的内容。
基于数据经更新8中内容。
调整属性，通过所有通道再次投放
8：进入其他需要用户数据组织内容的推荐通道
8：收集内容分类Badcase

内容Rank

- 各细分类别内部Rank
- Rank考虑：
- 前期：Cluster Score，内容质量（长度、标题长度、实体词密度、图片数量、图片质量）、Page Rank、Site Rank、Media Rank。
- 后期：以上加发布时长、点击率衰减（文章实效评价）、转化率分布（识别三俗内容）、用户行为转化率（时间衰减，增加或减少内容TTL）、次数转化率分布（疲劳阅读）、阅读时长与下拉频次。

用户建模

与内容分类对应

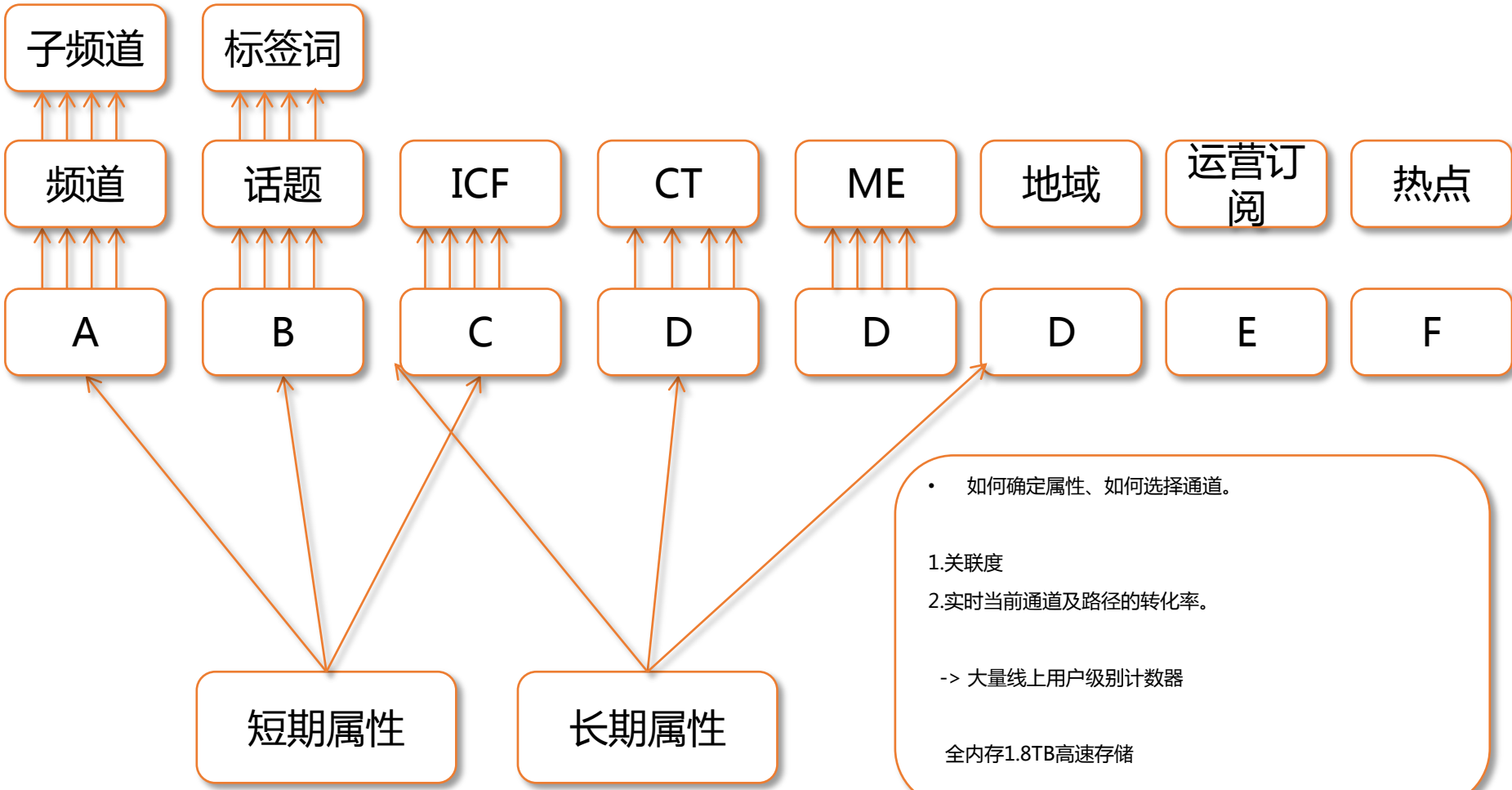
分为长期短期两套体系

长期：用户半年阅读行为、更新周期3天

短期：用户最近两天阅读行为、更新周期10秒。

用户阅读历史

投递逻辑



- 如何确定属性、如何选择通道。

1.关联度

2.实时当前通道及路径的转化率。

-> 大量线上用户级别计数器

全内存1.8TB高速存储

投递逻辑:通道转化率计算

用户属性

用户（2亿）

x

通道属性

推荐引擎（8+）

标签词（6万）

主题 （5000）

频道子频道（70）

=

上百亿计数器

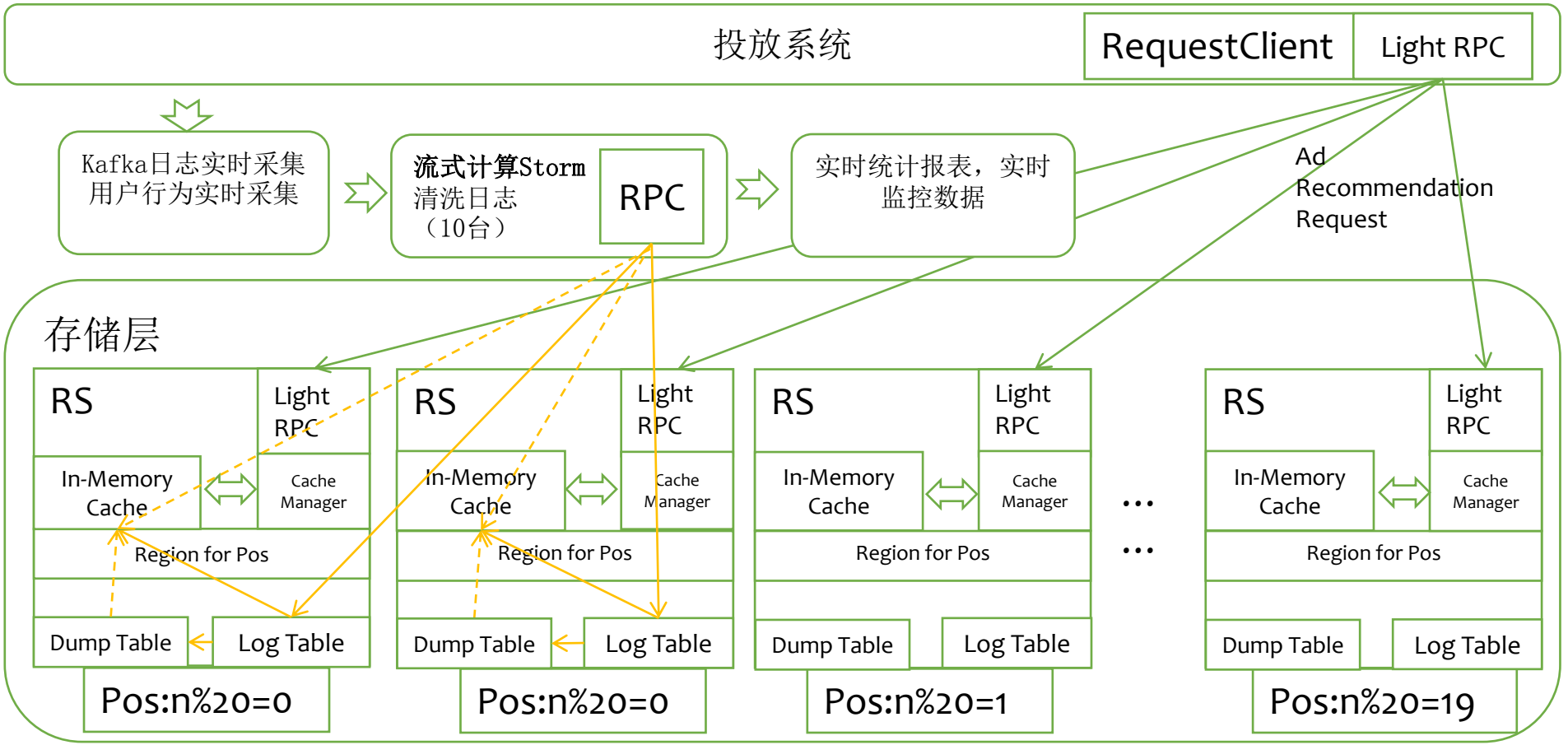
每秒6000*6*20次、

百万次库操作

约1.8TB*2内存数据库

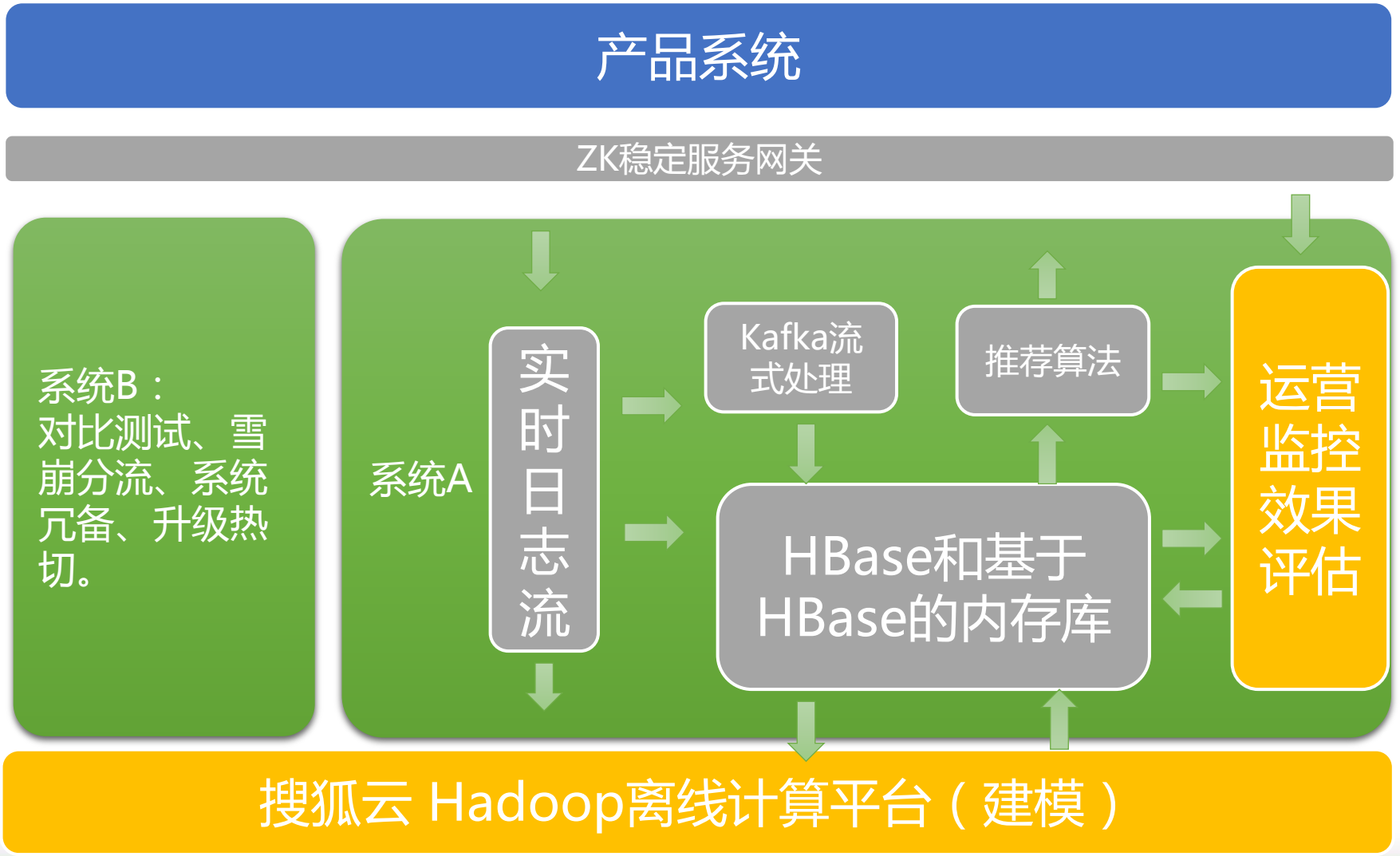
- 用户历史（用户*30*15）
- 每个存储48小时及最近15分钟点击曝光
- 系统容量每秒5000次请求，每日

MemT基于HBase的内容存储系统



数据冗余，压缩层次

系统框架


产品系统
ZK稳定服务网关
系统B：
对比测试、雪崩分流、系统冗备、升级热切。
系统A
实时日志流
Kafka流式处理
推荐算法
HBase和基于HBase的内存库
运营监控效果评估
搜狐云 Hadoop离线计算平台（建模）

评价体系

线上系统评价内容：推荐总量PV、用户使用次数、转化率、超时

考评来源：自媒体、机构媒体、第三方合作

考评频道：生活、娱乐、科技、财经

考评引擎：频道子频道、主题、协同、地域

灰度系统评价内容：人均消费次数、人均阅读时长、人均消费数量、人均分享评论、推荐转化率

验证：新算法、新内容效果

问题难点

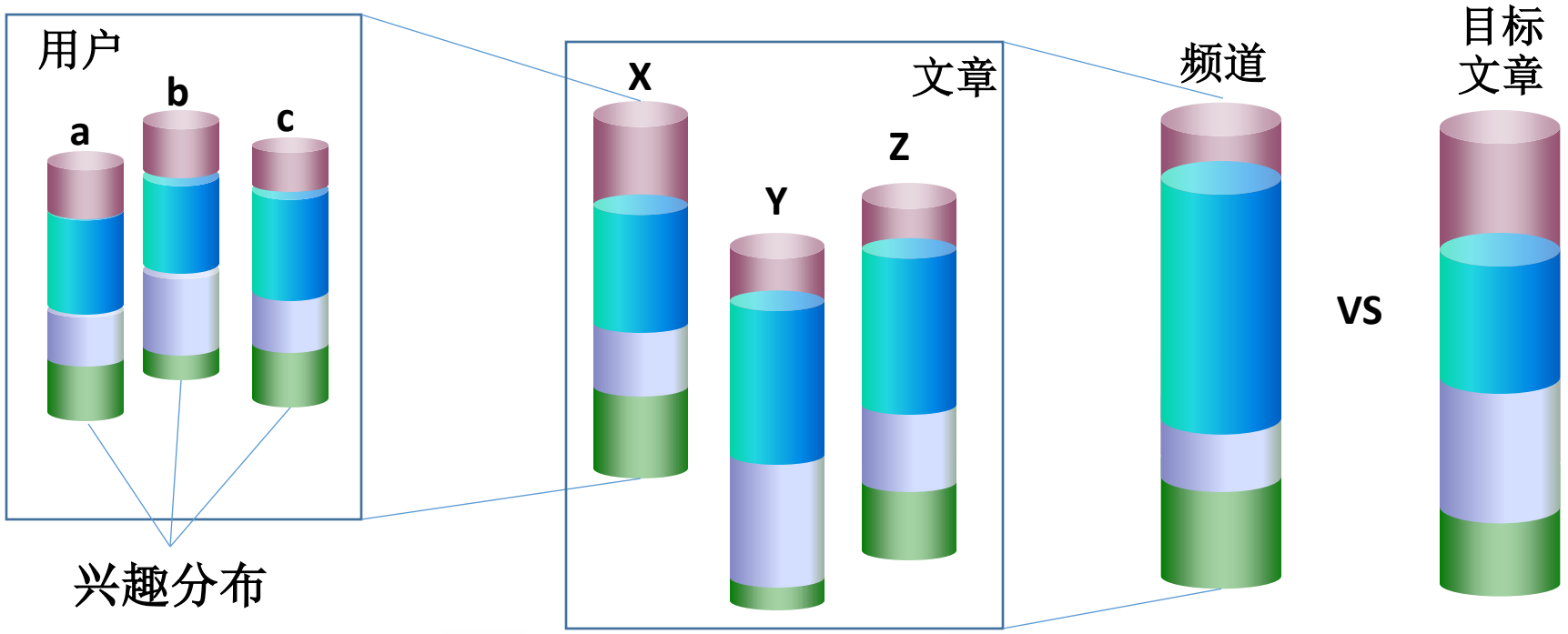
- 数据稀疏、内容冷启动、用户冷启动
- 噪音处理：三俗内容

噪音处理：三俗内容

- **三俗内容对点击转化率有很大提升。18+%**
  - 转化提高伴随曝光总量PV下降。
    - 转化提高，用户阅读时间提高，减少刷新次数？
    - 用户粘性下降，减少刷新次数？
  - 平均阅读时间下降，刷新次数下降，打开次数下降，用户粘性下降。
- **三俗内容危害**
  - 扰乱用户兴趣
  - 扰乱内容Rank

噪音处理：三俗内容

- 三俗：共性也特性
  - 低质（“三俗”）新闻过滤、制止热点内容



噪音处理：三俗内容

- **其它方法：**
  - 细化分类拆解内容：生活、文化频道内容三俗比率高。近一步细分出电影、旅游、读书、美食等频道，减少三俗内容干扰。
  - 提高阅读时间、分享收藏权重Rank权重。
  - 人工、根据关键词密度机器封停

- **效果：**
  - **转化率下降、推荐总量上升、用户打开次数使用时间增加。**

内容垂直度得分

科技频道、生活频道top100新闻垂直程度得分。

厦门"最赞APP"榜单出炉厦App属于国内一线阵营	12.40676
裁员3000人传eBay为分拆PayPal酝酿最大裁员	9.013341
Apple联合IBM对外发放10款企业级用户的APP	8.631804
女星比“纯”,拿什么来拯救你我的初恋!	7.705627
网易、陌陌撕逼大战,又被玩成群体狂欢	6.544471
10张图告诉你华为如何从迷茫中走出来,成长为一家年收入389亿美金的公司	5.97364
海盗湾遭查封创始人首发声希望就这么关下去	4.942631
韩国在华创业者眼中的陌陌上市可能是一个灾难	3.996687
魅蓝狙击红米小米威武不再	2.970068
镁电池或将成为下一代新能源电池的突破点	1.938068
雷军奋不顾身玩智能家居,小米手机若拜拜了雷总拿啥连接一切?	0.848008
幸福女人生命中必不可少的七个男人	0.771584
扔掉家里这八样东西!	0.692403
国家工商总局：天猫京东等电商双十一仍存售假	0.597741
不得不说,那些简直想给自己两巴掌的装修经历	0.499899
有关于手机电池的三大真相你知道几个?	0.369147
婚前处女与不如婚后贞女	0.284278
婚恋心理男人不娶性感女人5大原因	0.26716
揭秘传闻中9大神秘生物真相,绝对惊爆你的眼球	0.239115
一大波笑翻GIF来袭~二货老爸!	0.234326
厨房灶台朝向与风水,灶台装修不得不看的风水	0.233237
吃货就是吃货,笑死也值了	0.23213
这竟是19世纪男性的忠贞裤！古人也是脑洞大开啊！	0.223196
未来中国发生巨变内幕,一条消息震撼亿万国人	0.216264
中国学渣逆袭娶乌克兰女神晒幸福	0.209791
怎么让鸡汤的鸡味更浓？7个小窍门搞定	0.171698
笑的我肚皮都抽筋了,必须分享	0.163005
笑死人不偿命,惊呆了!!	0.100011

liujiaplus@163.com weixin: liujiaatict

用户冷启动和内容冷启动

- 用户冷启动
  - 关联用户信息
    - 用户挖掘挖掘其他属性利用sohu passport
    - 关联多个设备和账户
    - 获取微博客数据
  - 短期兴趣快速拟合
    - 快速反馈、成短期兴趣及各通道推荐转换率数据。
- 内容冷启动
  - 提高试投放PV利用效率，垂直内容找到专家用户评判质量。
    - 内容去重聚合提高数据覆盖
    - 强化横向纵向关联关系
    - 强化Topic分类精度
    - 频道中引入编辑内容
    - 引入搜狗PV

BDTC
2014 中国大数据技术大会
BIG DATA TECHNOLOGY CONFERENCE
暨第二届CCF大数据学术会议

感谢！ Q&A