新浪微博

基于微博用户关系和行为的用户建模

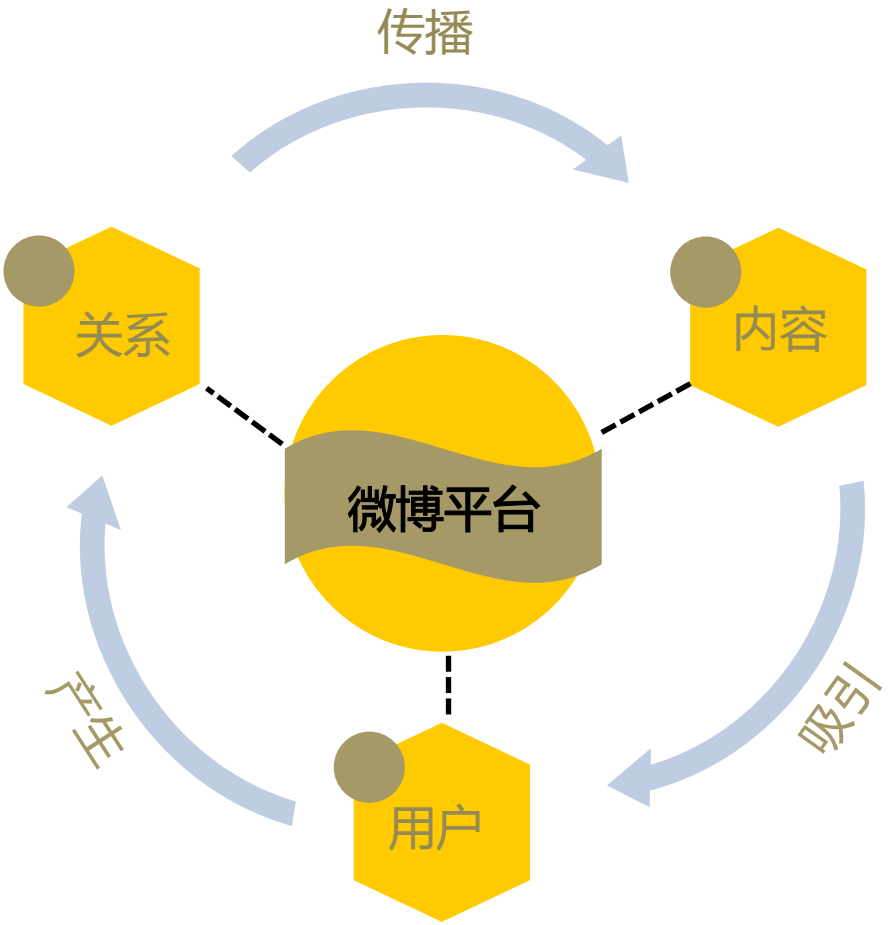
新浪微博大数据-朱红垒
微博：@叠石

提纲

- 微博及大数据
- 大数据标签体系
- 用户能力标签
- 用户兴趣标签

微博及大数据

中国最大的社交媒体平台



微博沉淀了海量的用户、关系、内容、行为数据

微博及大数据

- 用户
  - 注册人数：10亿
  - 月活人数：1.98亿
  - 日活人数：8900万
- 关系：
  - 关注关系：近千亿
  - 分组关系：50亿+
- 内容
  - 日增博文：1亿+
  - 日增原创：4000万
- 行为
  - 转发：6000万
  - 评论：3000万
  - 赞：1亿
  - 收藏：1000万
  - 查看：200亿

微博及大数据

微博大数据要做什么

帮助用户发现感兴趣的内容

加快有价值内容的传播效率

目标如何实现

挖掘有能力生产垂直领域优质内容的用户

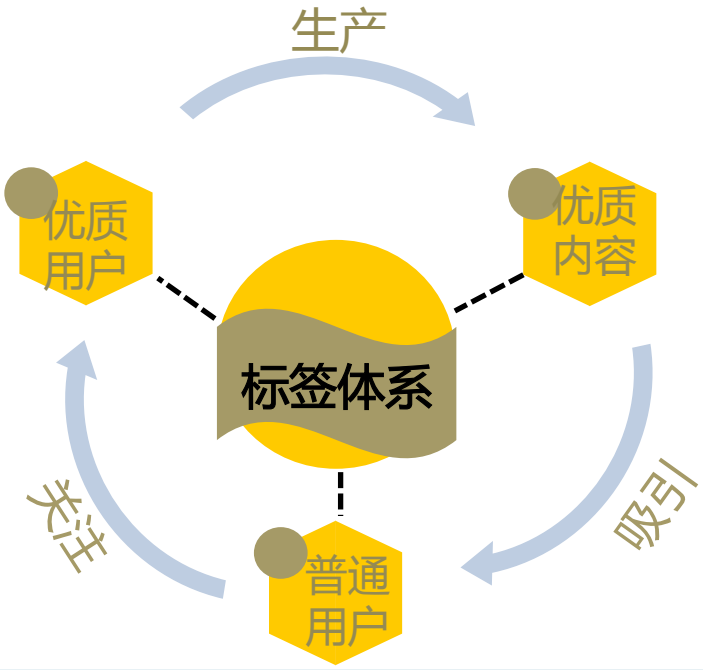
挖掘用户内容消费的兴趣偏好

工作如何串联

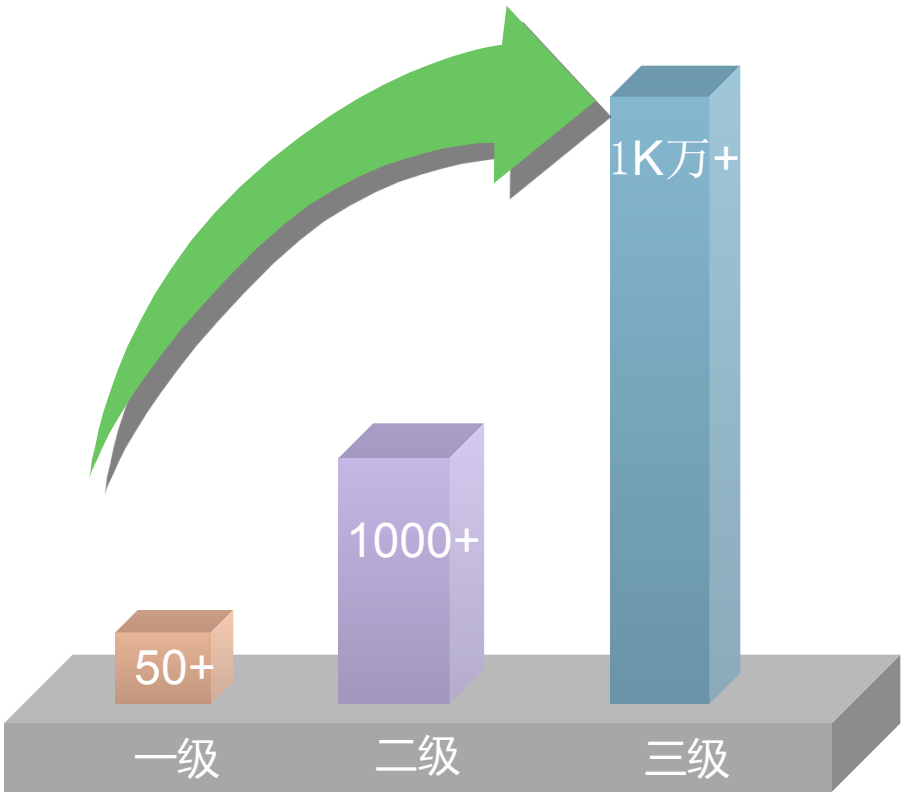
用户能力标签

用户兴趣标签

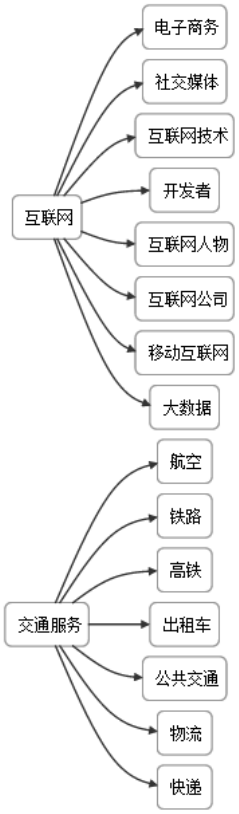
微博内容标签

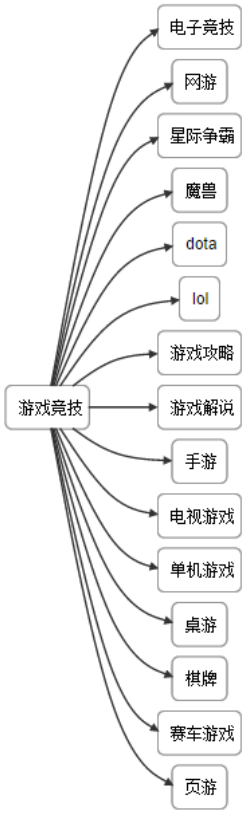


大数据标签体系



标签规模





三级标签举例：php、皇马、纸牌屋等

用户能力标签—产品形态

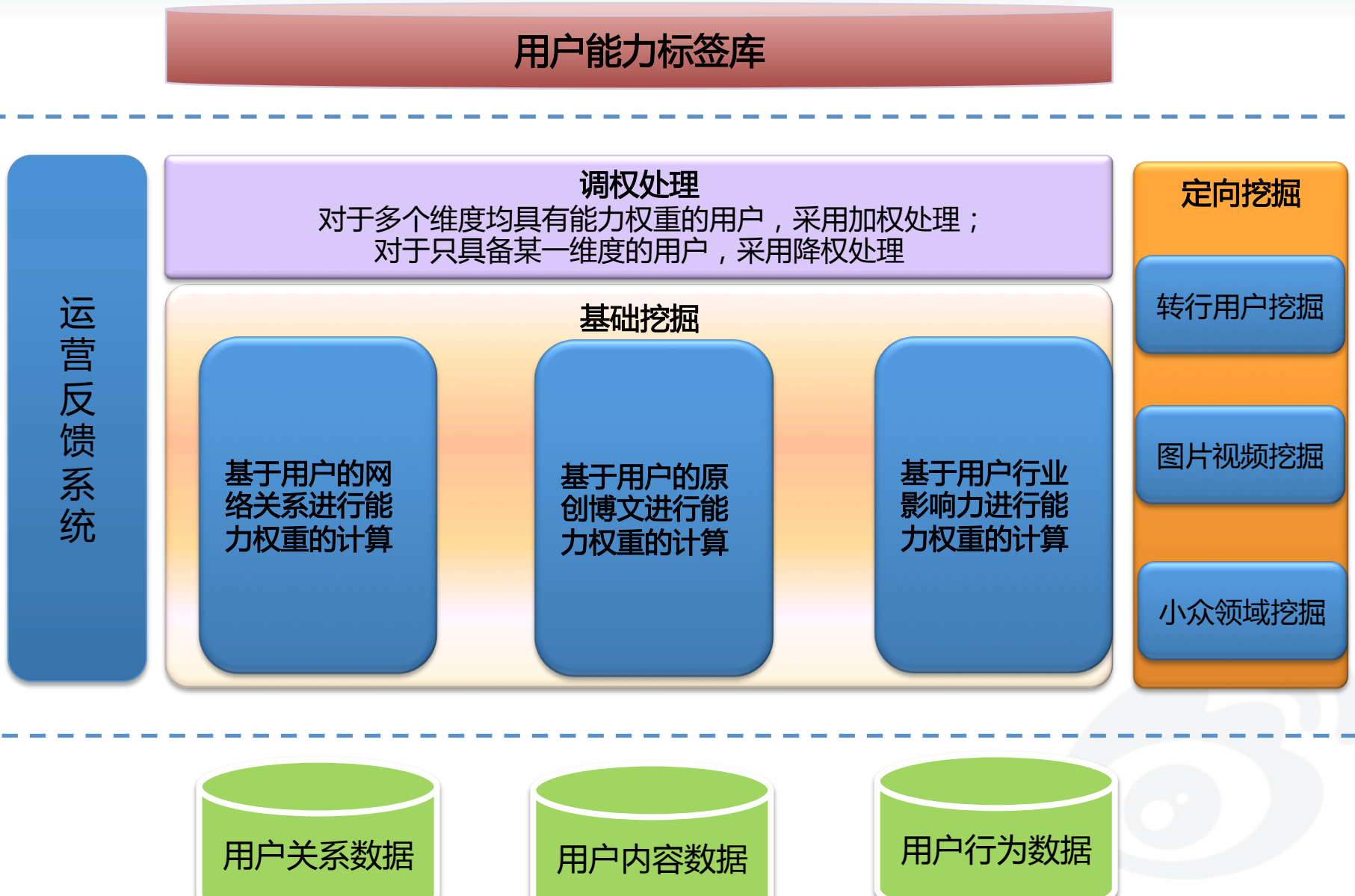
微博找人

直接推荐各行各业的能力用户

微博头条

输出各领域原始语料的专家库

用户能力标签—整体框架


用户能力标签库
运营反馈系统
调权处理
对于多个维度均具有能力权重的用户，采用加权处理；
对于只具备某一维度的用户，采用降权处理
基础挖掘
基于用户的网络关系进行能力权重的计算
基于用户的原创博文进行能力权重的计算
基于用户行业影响力进行能力权重的计算
定向挖掘
转行用户挖掘
图片视频挖掘
小众领域挖掘
用户关系数据
用户内容数据
用户行为数据

用户能力标签— 策略算法

(1) 基于决策树的分组名分类算法：将分组名分为兴趣分组名和熟人关系分组名

兴趣分组名用于用户的能力兴趣计算

熟人关系分组名用于用户的自然属性挖掘

(2) 基于用户关注关系数据计算用户在关系方面的能力：

利用兴趣分组名称构建出标签的相关词库，进而通过归一化公式计算出基础权重

通过认证信息、自标签进行权重调权，输出用户在关注关系方面的能力权重

(3) 基于用户发布内容数据计算用户在内容方面的能力：

用户在某个领域发布博文数量、纯度、互动量越高，在这个领域内容生产能力越大

$$S=\sum(f\times\alpha+c\times\beta+l\times\gamma)\exp(-\epsilon\times\Delta dt)$$

$$w=\frac{a}{1+e^{-\theta\times s}}-b$$

其中f为转发数,c为评论数,l为赞数, Δdt为博文发表时长

α,β,γ, θ, a,b为相关参数

(4) 通过PageRank计算用户在垂直行业的影响力：

通过PageRank计算具有一定内容生产能力和关系能力的用户群中每个用户的影响力

(5) 通过线性加权将用户的关系、内容和行业影响力计算为在这个垂直领域的综合能力：

用户能力标签归一化到0~100的区间，达到横纵向可比较

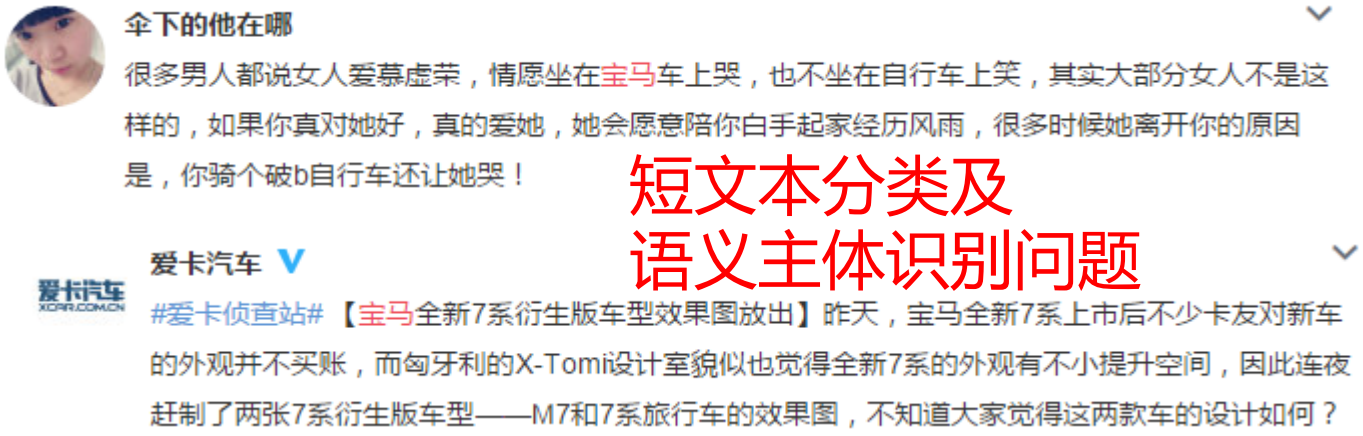
用户能力标签—主要问题

1. 标签的自动聚合及筛选

噪音问题

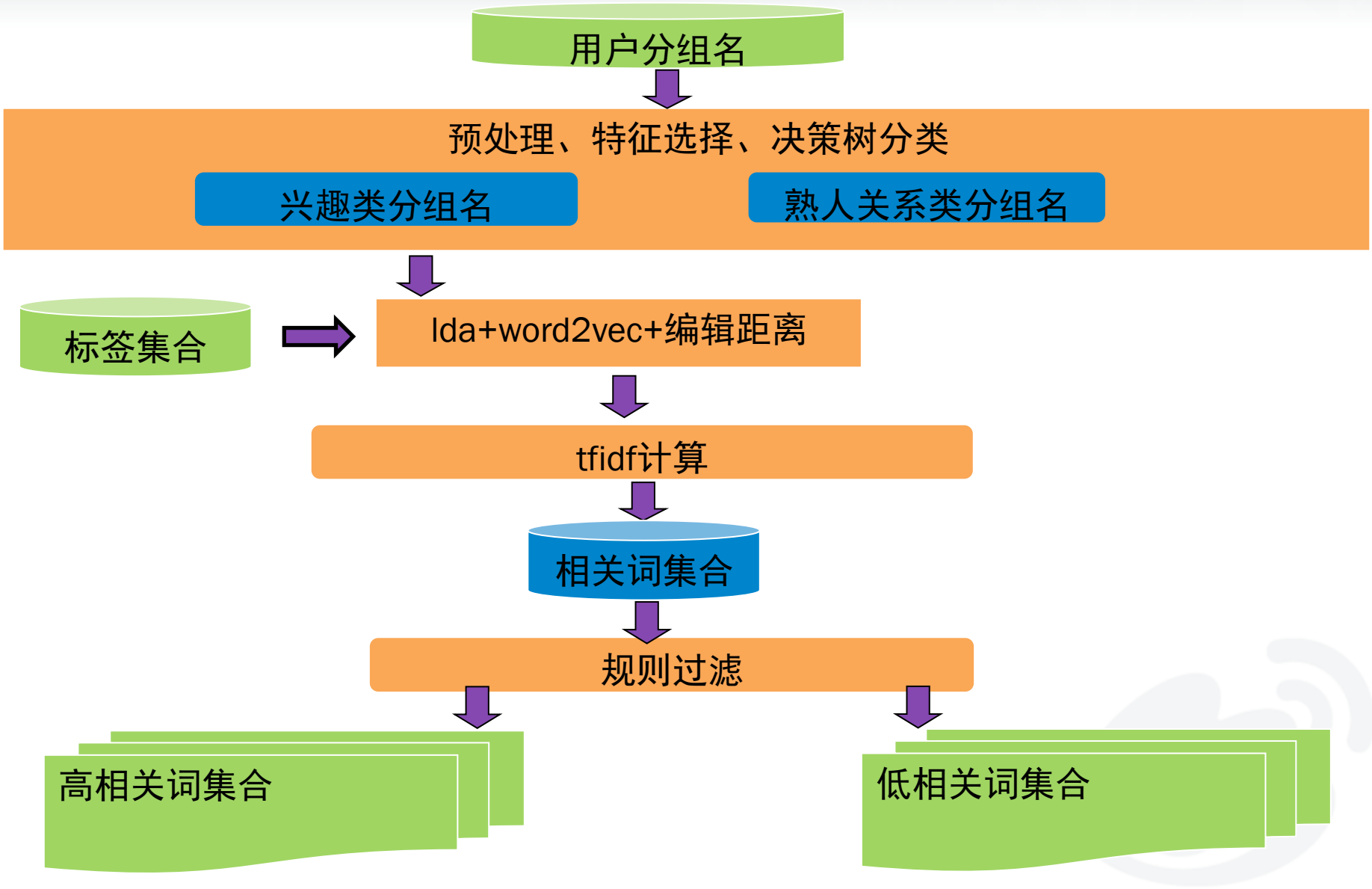
```
任志强  名人:407268  商业:146870  明星:126605  社会:58166  房地产:34959  经济
:29893  财经:15168  专家:14619  时事:11168  地产:5674  房产:4402  新闻:426
3  媒体:3853  公众人物:2577  公知:2180  资讯:1708  学者:1694  生活:153
3  金融:1317  投资:1094  偶像:865  时政:795  财经名人:641  互联网:6
14  政治:598  媒体人:586  老师:557  建筑:555  达人:555  地产名人:550
   地产圈:522  大佬:511  评论:507  传媒:482  艺人:480  财经人物:445
   boss:401  经济名人:391  经济学家:382  地产界:373  股票:371  时评
:337  网络红人:333  知识:332  理财:315  名人名言:296  finance:296
地产商:292  任志强:286  明人:282  娱乐圈:281  开发商:275  娱乐明星:273
```

2. 微博短文本的识别问题



短文本分类及
语义主体识别问题

用户能力标签—标签自动聚合流程



用户能力标签—效果

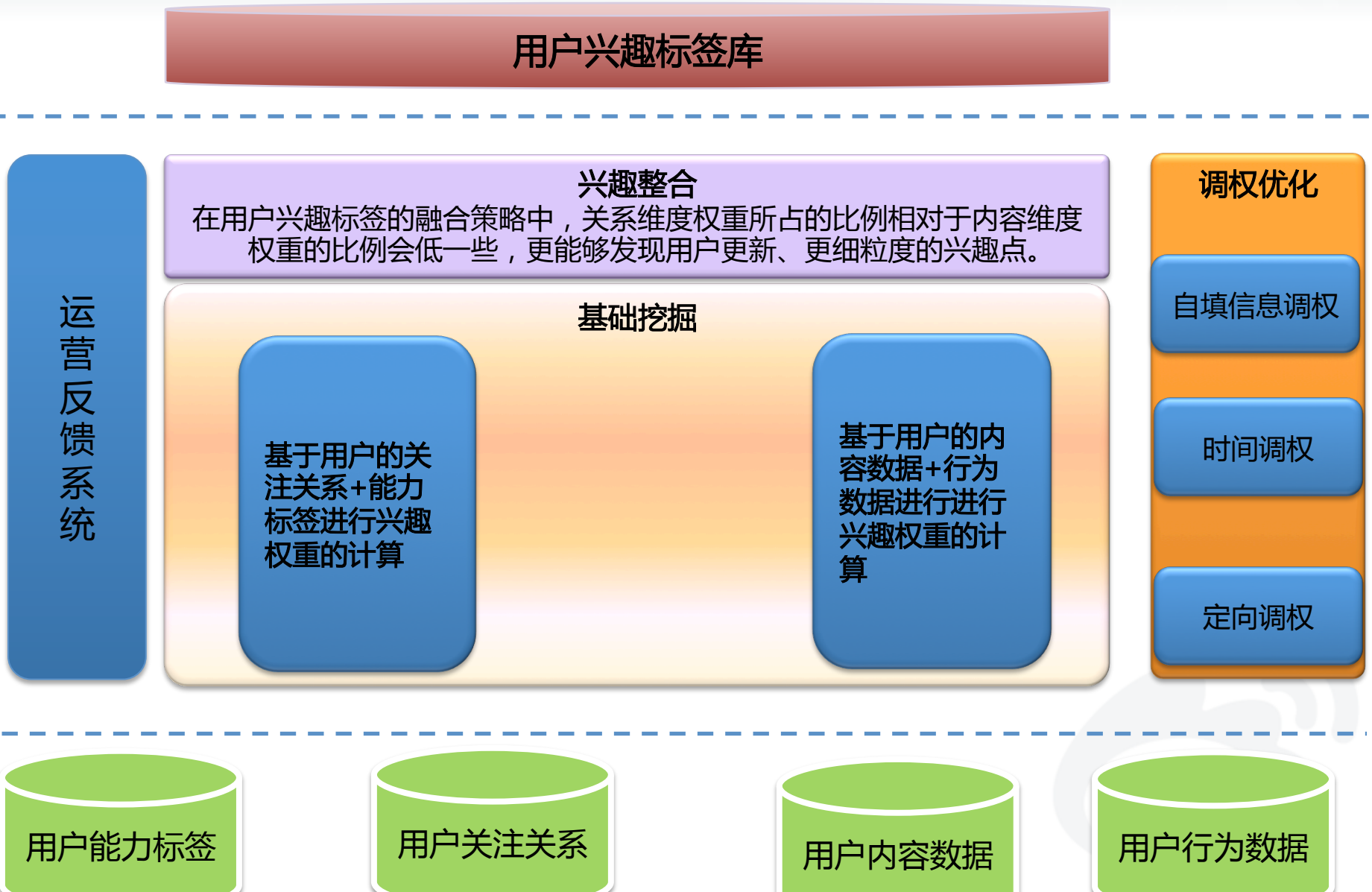
- 挖掘出120万能力用户，覆盖月活粉丝1.6亿
- 微博用户中娱乐、互联网、财经行业名人最多
- 微博用户中动漫、美食、旅行行业精英最多
- 微博聚集了近万名互联网技术牛人

用户兴趣标签—产品形态

基于兴趣推荐用户感兴趣的文章

基于兴趣Push用户一段时间内关注人发的但是没看过的微博

用户兴趣标签—整体框架


用户兴趣标签库
运营反馈系统
兴趣整合
在用户兴趣标签的融合策略中，关系维度权重所占的比例相对于内容维度权重的比例会低一些，更能够发现用户更新、更细粒度的兴趣点。
基础挖掘
基于用户的关注关系+能力标签进行兴趣权重的计算
基于用户的内容数据+行为数据进行进行兴趣权重的计算
调权优化
自填信息调权
时间调权
定向调权
用户能力标签
用户关注关系
用户内容数据
用户行为数据

用户兴趣标签—策略算法

(1) 圈定各垂直领域的能力用户集合:

根据用户能力标签分数分布以及各分数段的意义，圈定垂直领域的能力用户集合

(2) 根据用户对能力用户的关注关系计算用户在关系方面的兴趣:

关系兴趣权重的计算规则：根据w1和w2最终确定关系兴趣的权重

$$w1 = \frac{a}{1 + c^{-\theta}} - \mathrm{b}$$

其中c为关注某一类能力用户的数量
a,b, θ 为相关参数

$$w2 = \frac{e}{1 + \mathrm{d}^{-\varepsilon}} - \mathrm{f}$$

其中d为关注某一类能力用户的数量与总关注数的比值
e,f, ε 为相关参数

(3) 根据用户对内容产生的行为计算用户在内容消费方面的兴趣:

微博行为包括：原创，转发，评论，赞，收藏，查看微博等十几种行为

不同的行为对应不同的分值，最终通过归一化公式计算用户消费内容的兴趣权重

$$w3 = \frac{a}{1 + \sum_{i=1}^{n} \alpha_i c_i^{\ -\theta}} - \mathrm{b}$$

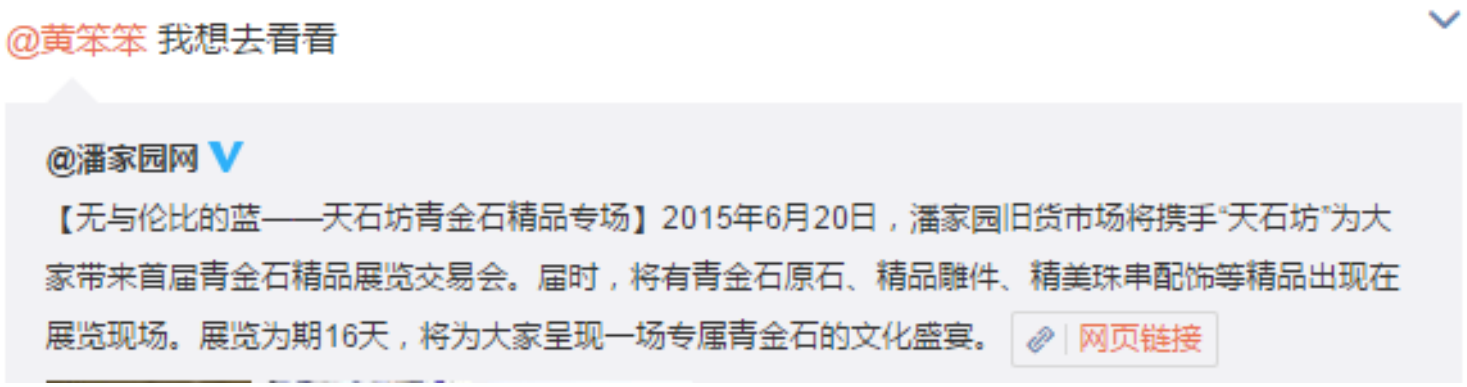
其中 α_i 为第i项行为的权重, c_i 为第i项行为的得分，a,b, θ 为相关参数

(4) 通过线性加权计算用户的综合兴趣调权:

通过不断的迭代测试，用户在内容消费方面的权重更高一些

用户兴趣标签— 主要问题

1. 用户的兴趣相对于能力而言是时间敏感的，如何在用户的兴趣权重上体现出时间敏感性是一个关键问题



2. 所有用户都对实时，旅游，明星感兴趣?

3. 活跃用户体现出的兴趣极为广泛

用户兴趣标签—用户行为权重的时间衰减

基于时间维度的行为热度衰减：在博文消费方面，用户通过转发、评论、赞等行为来表达自己对领域博文的兴趣，其表达的兴趣热度及重要性随着时间是逐步衰减的，我们通过牛顿冷却定律来量化衰减的程度。

$$c = c_0 \, e^{-\alpha(t-t_0)}$$

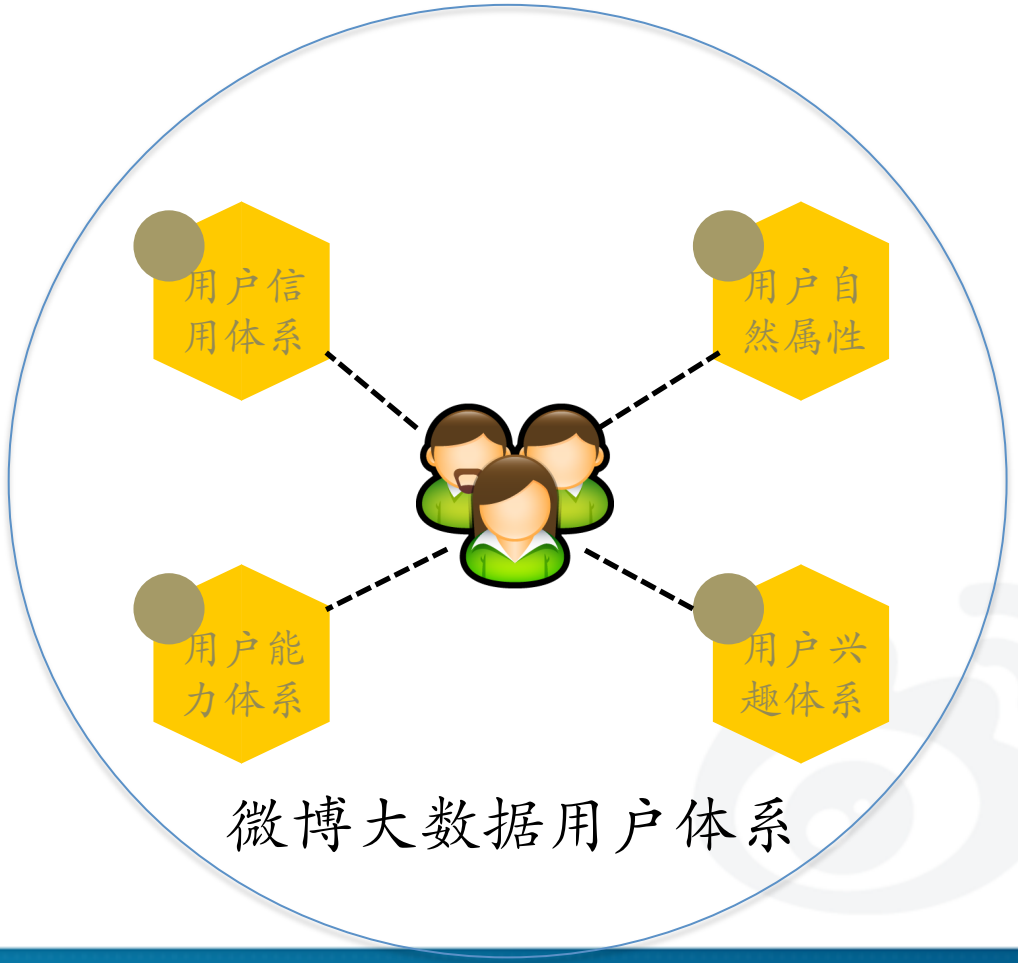
其中
c为当前兴趣权重，
c_0为初始兴趣权重，
α为冷却系数，
$\Delta dt = t - t_0$为间隔的天数

用户兴趣标签—效果

- 挖掘出1.6亿用户的精准兴趣，覆盖微博月活75%
- 微博用户中对娱乐、时事、互联网感兴趣的人最多
- 微博聚集了110万对互联网技术感兴趣的人

用户标签的规划

- 用户身份
- 用户即时兴趣
- 用户质量等级



新浪微博

Thanks & QA

weibo.com