网易微博推荐系统
—之内容推荐篇

高金楠

2013.01

记录我们的微生活

网易微博
t.163.com

目 录

contents

推荐算法分类

常用的推荐算法：

- 协同过滤的推荐
- 基于内容的推荐

  该方法是对协同过滤的延续和发展，主要借鉴了信息抽取和信息过滤的研究成果，依据被推荐项目的内容特征来进行推荐。
- 基于人口统计学的推荐
- 混合的推荐

  以上几种方法的混合。

内容推荐步骤

1. **内容抽取：**为每个item抽取出一些“内容”（也就是item的特征）来表示此item。

2. **兴趣学习：**利用一个用户过去喜欢(及不喜欢）的item的特征数据，来学习出此用户的喜好特征。

3. **生成推荐：**比较用户的兴趣与候选item的特征，为此用户推荐一组相关性最大的item。

优点：

1. 个性化推荐
2. 新增的item立刻被推荐

不足：

1. item内容抽取困难
2. 用户兴趣学习具有局限性
3. 不能为新用户做出推荐

微博内容推荐工作

发现：向用户推荐各分类下达人的优质微博。非个性化

微博内容推荐工作

榜单：热门微博推荐。

热门转发 | 每日精选 | 热门专栏 | 精选段子 | 热门投票

张鸣：既然说已经废除了干部终身制，为何退休的人，再出来还要排名？退休之后，不就是平民百姓了吗？如果还排名，还享受特权待遇，那么，所谓废除干部终身制，就是假的。

+关注

54分钟前 来自享拍微博通

转发(121) | 收藏 | 评论(42)

任志强：用什么方法建立牢固的笼子，并把权力关进笼子？民主也许不是最好的，但却可能是唯一的方式。

+关注

昨天22:03 来自网易微博

转发(314) | 收藏 | 评论(93)

茅于轼：民主和专政是两个绝然相反的词。但是在我国宪法中居然把二者结合起来，称我国的政体是"人民民主专政"。最奥妙的是在民主和专政的前面加上"人民"两个字，就把最不合理的东西合理化了。什么难解释的东西前面一加了"人民"就变好了。人民警察，人民法院，人民政府，人民币，人民代表，人民政协……

+关注

昨天12:22 来自网易微博

转发(793) | 收藏 | 评论(346)

李子暘：再次重申：计划生育政策不是现在才变得不正确，才变得不合时宜，才变得有危害的。计划生育从出现的第一天起就是错误的，就是有危害的。人口数量，从来不需要政府控制，任何时候也不需要。中国过去的贫穷，和人口多一点关系也没有，完全是计划经济的结果。

+关注

昨天09:58 来自玛撒网

转发(960) | 收藏 | 评论(331)

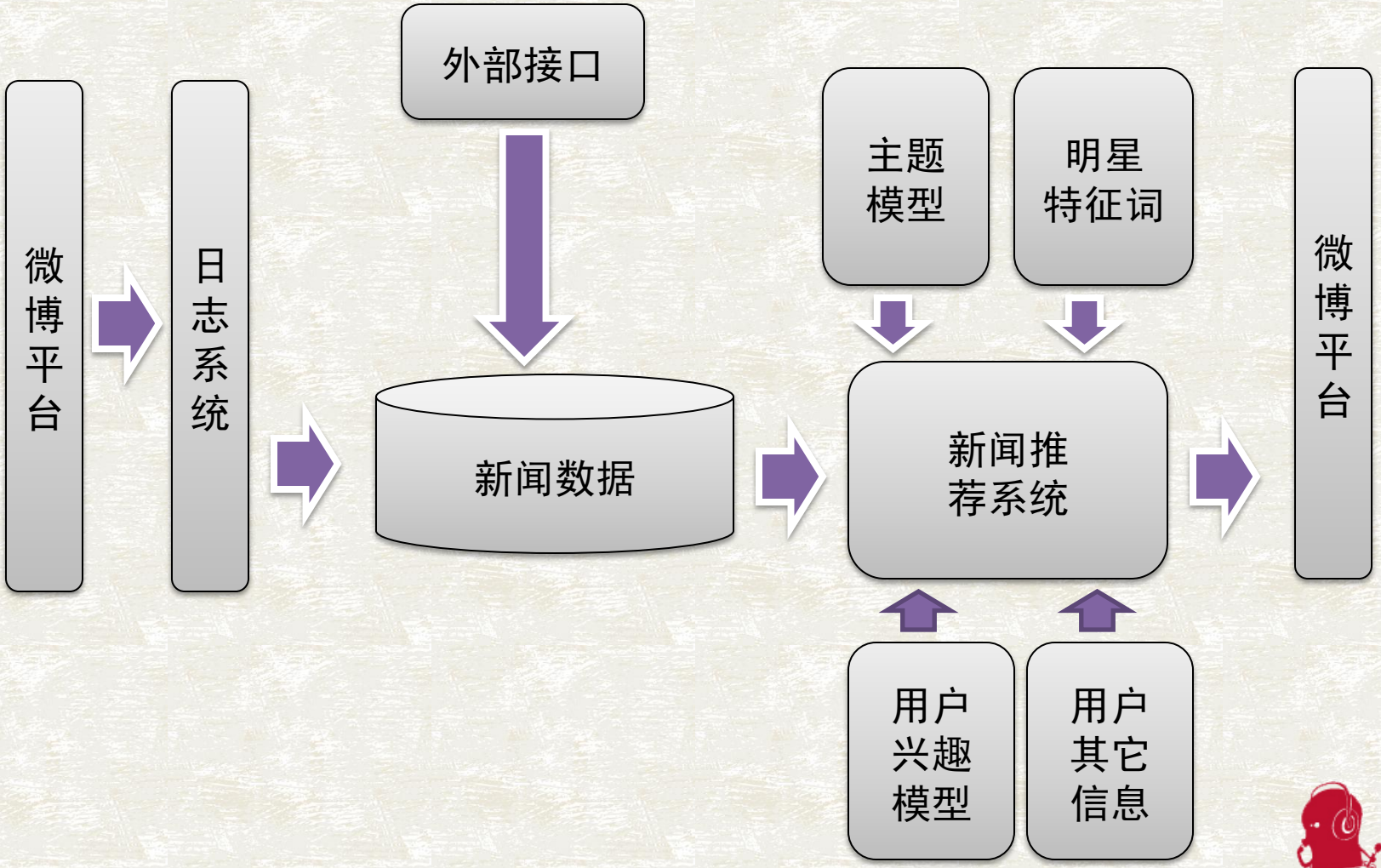
微博内容推荐工作

新闻推荐：找出新闻中涉及到的明星，将该新闻推荐给此明星的微博粉丝。

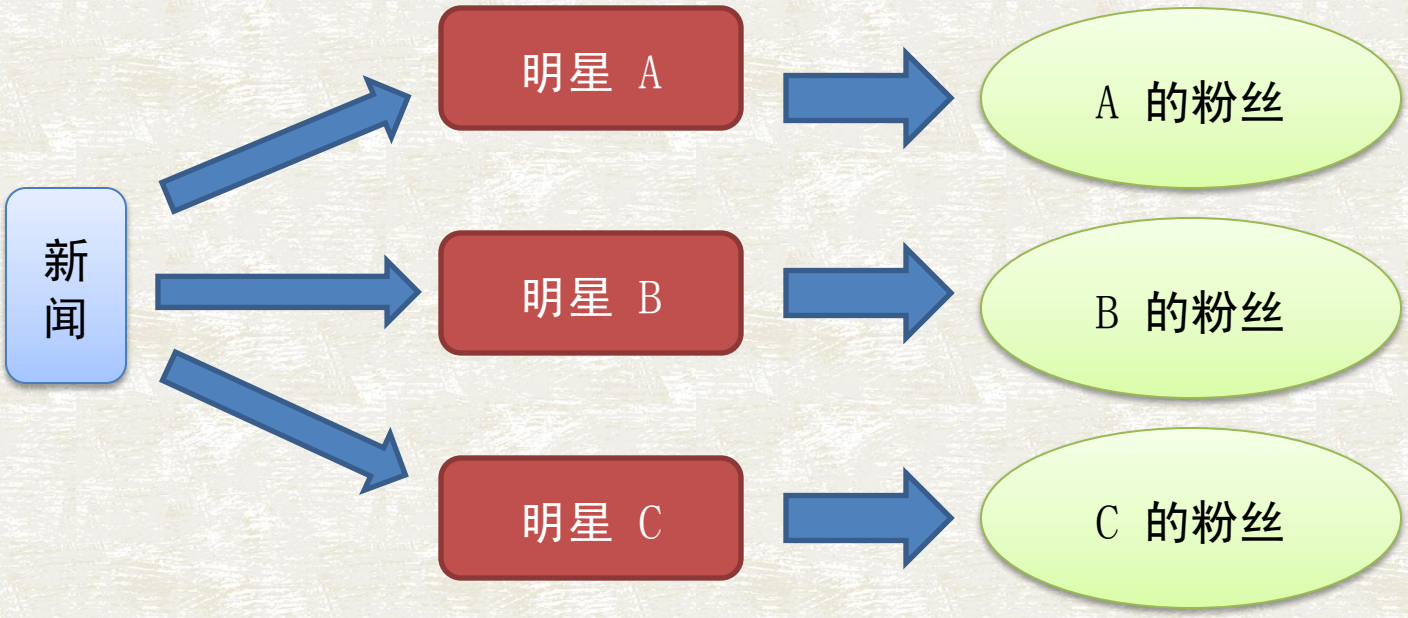
新闻推荐背景

网易新闻的官方微博，每天会转载大量的新闻至微博系统（5000/d）。帮助用户从大量的信息中发掘用户可能感兴趣的新闻。

新闻推荐系统框架



新闻推荐系统



新闻到明星的映射

- 主题分布的相似
  前提：一个人在一段时间内涉及的主题是一定的。

- 特征词过滤
  1. 明星的姓名
  前提：在一篇新闻中，只有出现了某个人的名字，这篇新闻才可能和这个人相关。
  2. 明星的特征词
  前提：对每个明星都存在一组具有识别性的特征词，当这组词中的一个或几个与其姓名同时出现时，这篇新闻与此人相关度会比较大。
  （如：{冯小刚 | 天下无贼，非诚勿扰，徐帆…} ）

主题模型

主题模型 （Topic Model）

主题模型，就是对文字中隐含主题的一种建模方法。在主题模型中，主题变现为词汇表上词语的条件概率分布。

例如：

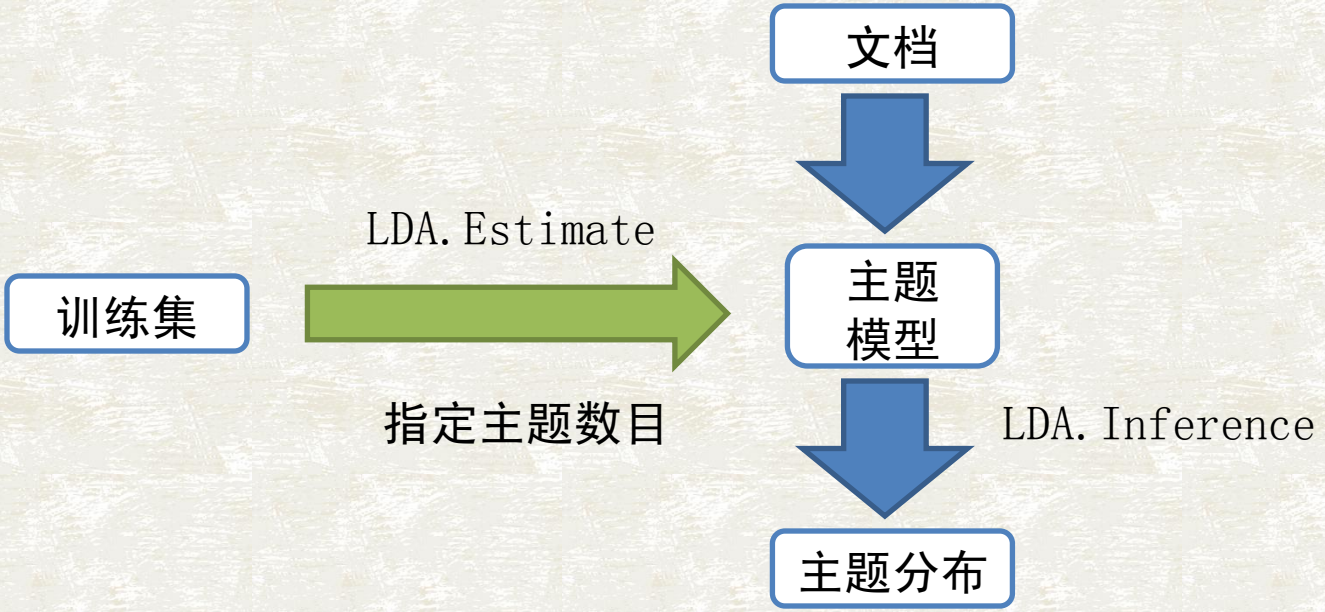
开出 号码 分析 推荐 上期 看好 和值 出号 走势 奖号 …

球员 拜仁 位置 球队 名单 出场 巴里 前锋 主力 中场 …

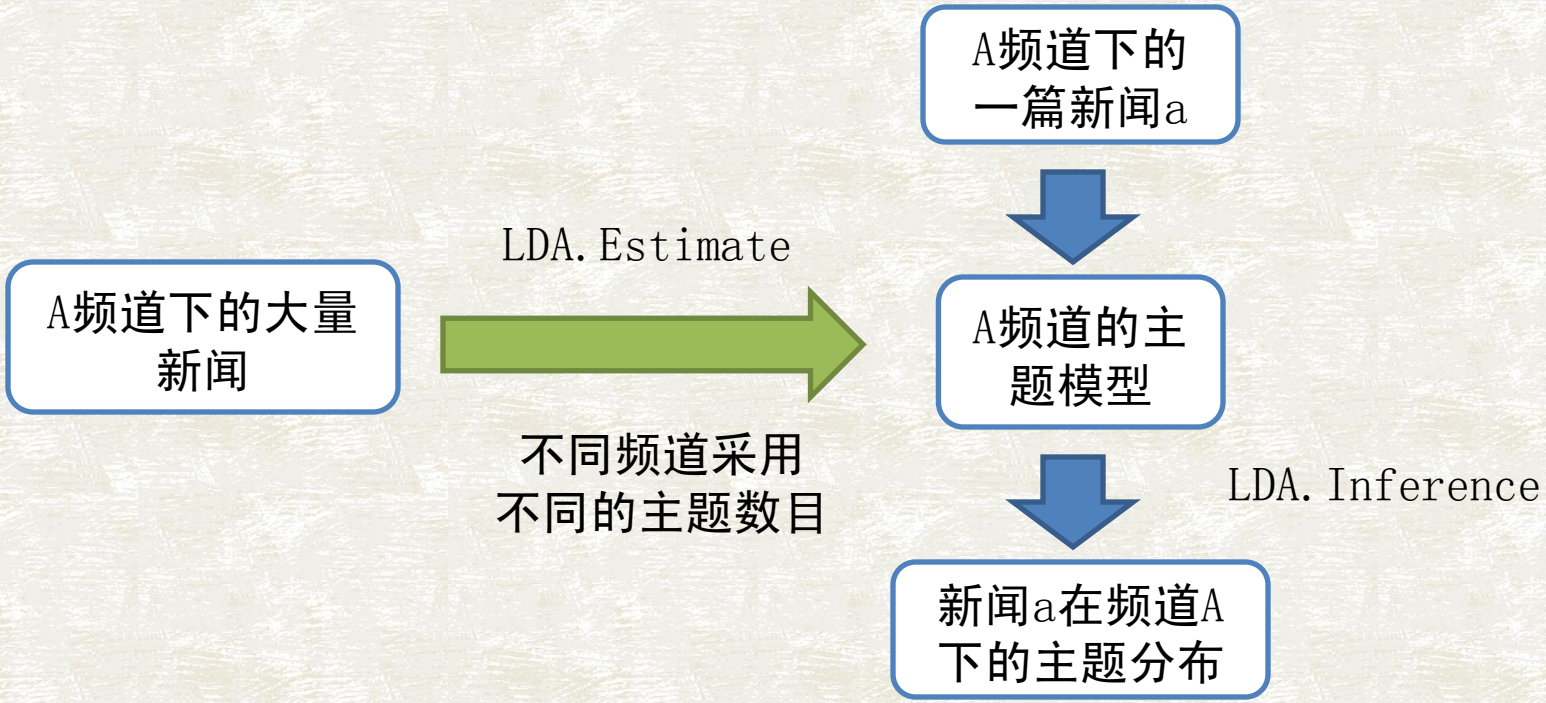
主题模型

算法选择 LDA

LDA 是一种非监督机器学习技术，可以用来识别大规模文档集或语料库中潜藏的主题信息。



新闻及明星的主题分布



明星在某频道下的话题分布 = Σ 明星在该频道下新闻的话题分布

主题分布的相似性

由**向量的夹角余弦**计算新闻与明星主题分布的相似度

新闻/话题	T 1	T 2	…	T n
News 1	P11	P12	…	P1n
News 2	P21	P22	…	P2n
News 3	P31	P32	…	P3n

明星/话题	T 1	T 2	…	T n
Star 1	P11	P12	…	P1n
Star 2	P21	P22	…	P2n
Star 3	P31	P32	…	P3n

$$sim(x, y) = \frac{\sum_{u=1}^{m} L_{ux} * L_{uy}}{\sqrt{\sum_{u=1}^{m} (L_{ux})^2 * \sum_{u=1}^{m} (L_{uy})^2}}$$

主题模型效果评测

精准度：返回相关文档占返回总文档的比例。

当我们不关心所有的返回结果时，通常只对排名最靠前的一部分结果，所以有时候只考察对前n条结果的评价。

测试结果

P5	0.73
P10	0.75
P20	0.75
P30	0.70

特征词提取

- 作品集

  在网络上抓取明星的相关作品，当做这些人的特征词。这部分特征词与相应该明星的相关性比较高。

- 利用tf-idf提取特征词

  针对明星的历史新闻，通过tf-idf算法提取出相关的特征词。

tf-idf提取特征词

在明星的历史新闻 集合上，利用tf-idf提取特征词。

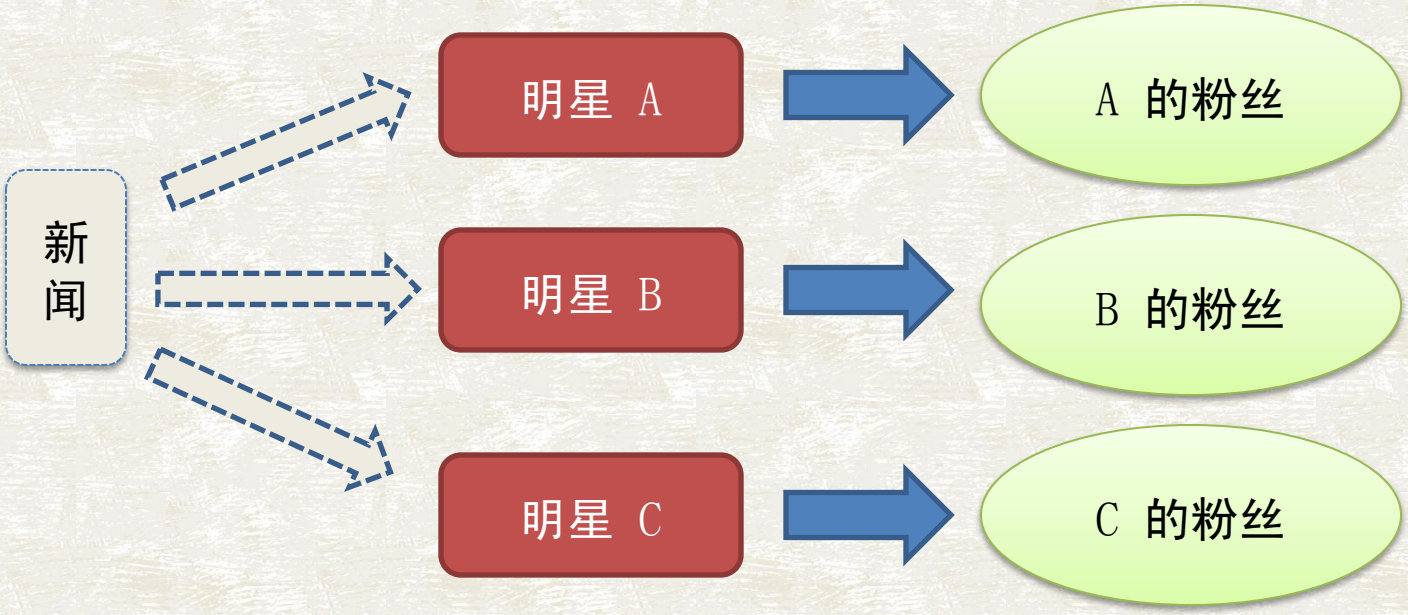
方案1.

将一个频道下每个明星的所有历史新闻合并成一个大文档，在大文档上做tf-idf，提取出每个人的特征词。

方案2.

把一个频道下所有明星的每篇新闻，看成一个文档， tf-idf提取出每篇新闻的特征词。再将一个明星所有新闻的特征词统计词频，求Top N。

新闻推荐系统



将新闻推荐给用户

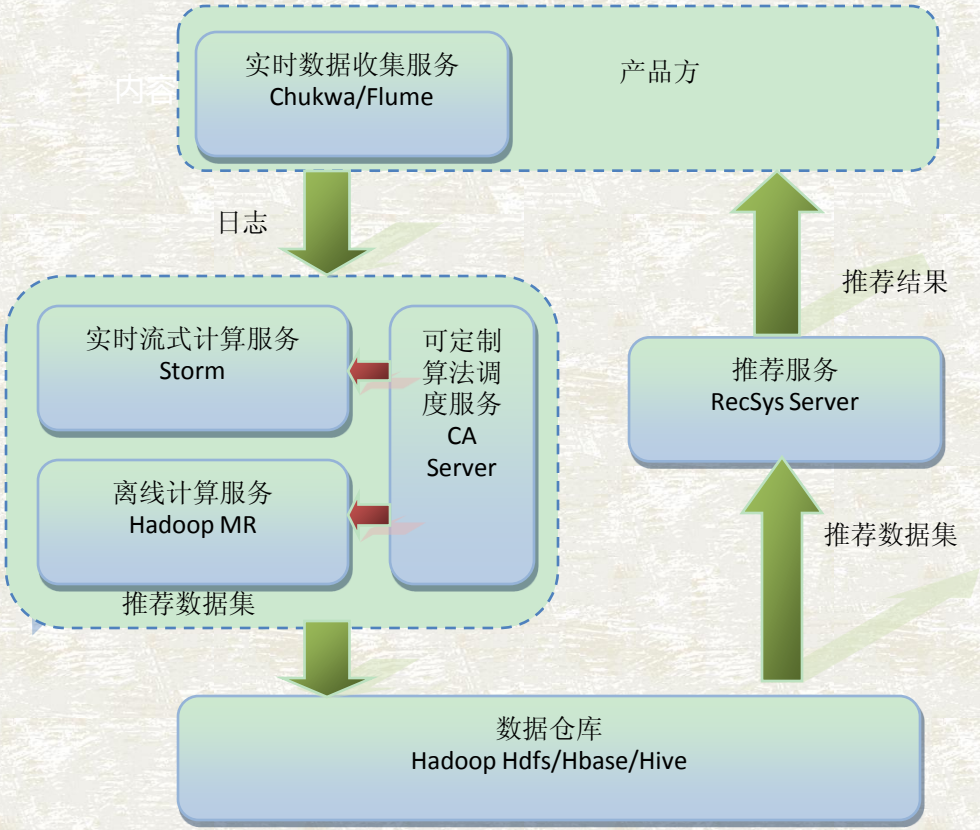
为用户提供个性化的新闻推荐

- 用户对兴趣的选择
- 用户与明星的互动
- 用户兴趣模型

新闻推荐优化方向

- 算法的优化
  - 分词时词性的标注
  - 计算并行化
  - tf-idf基于词语出现位置信息
- 用户行为的实时反馈
- 完善的推荐效果评估

推荐引擎


实时数据收集服务
Chukwa/Flume
产品方
日志
实时流式计算服务
Storm
离线计算服务
Hadoop MR
可定制
算法调
度服务
CA
Server
推荐数据集
推荐结果
推荐服务
RecSys Server
推荐数据集
数据仓库
Hadoop Hdfs/Hbase/Hive

Q&A

Thank You~~~

记录我们的微生活

网易微博
t.163.com