**推荐系统**

软件设计说明书

# 概述

## 需求

即时快捷地向用户推荐自己感兴趣的文章和话题

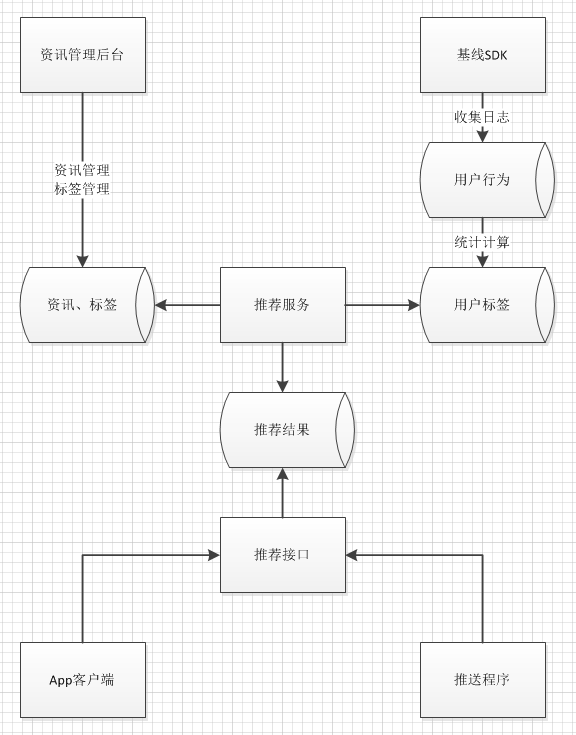
# 技术选型

推荐方法：基于内容特征的推荐

关键步骤：

1. 标签库（程序自动生成 + 人工编辑）
2. 给资讯打标签（程序自动提取 + 人工编辑）
3. 给用户打标签（用算法计算用户行为数据得到）
4. 把和用户标签相似度高的资讯推荐给用户

# 架构



# 详细架构

# 数据结构设计

## 资讯数据源

|  |  |  |  |
| --- | --- | --- | --- |
| **表名** | **t\_information** | **表类型** | **mysql** |
| **字段名称** | **类型** | **默认值** | **说明** |
| id | bigint |  | 资讯ID |
| category\_id | bigint |  | 类别ID |
| idwords | varvhar(255) |  | 标签ID:相关度 |
| publish\_time | datetime |  | 发布时间 |
| 其他 | mysql热备 | | |

## 资讯点击日汇总

|  |  |  |  |
| --- | --- | --- | --- |
| **表名** | **d\_info\_click** | **表类型** | **hive** |
| **字段名称** | **类型** | **默认值** | **说明** |
| info\_id | bigint |  | 资讯ID |
| user\_count | int |  | 用户数 |
| click\_date | string |  | 点击日期，分区字段 |
| 其他 | 数据来源：用户点击资讯日汇总表d\_user\_click  存储规则：按天分区  更新策略：每小时更新 | | |

## 用户点击资讯日汇总

|  |  |  |  |
| --- | --- | --- | --- |
| **表名** | **d\_user\_click** | **表类型** | **hive** |
| **字段名称** | **类型** | **默认值** | **说明** |
| device\_id | varchar(50) |  | 设备ID |
| info\_id | bigint |  | 资讯ID |
| event\_name | varchar(50) |  | 事件名称 |
| click\_count | int |  | 点击次数 |
| click\_date | string |  | 点击日期，分区字段 |
| 其他 | 数据来源：用户点击事件表event  存储规则：按天分区  更新策略：每小时更新 | | |

## 用户标签

|  |  |  |  |
| --- | --- | --- | --- |
| **表名** | **user\_profile** | **表类型** | **hbase** |
| **主键、列族** | **字段名称** | **字段解释** | **说明** |
| rowkey |  | 主键 | device\_id |
| timestamp |  | 时间戳 | 写入时间 |
| tag |  | 列族 |  |
| tags | 标签 | 标签ID:标签相关度 |
| wtags | 权重标签 | 标签ID:标签权重 |
| 其他 | 数据来源：用户点击事件表event、资讯表t\_information  更新策略：每小时更新  计算过程：见[用户标签计算过程推导](#_用户标签计算过程推导) | | |

## 活跃用户

|  |  |  |  |
| --- | --- | --- | --- |
| **表名** | **active\_device** | **表类型** | **mysql** |
| **字段名称** | **类型** | **默认值** | **说明** |
| device\_id | varchar(50) |  | 设备ID |
| active\_time | datetime |  | 活跃时间 |
| 其他 | 数据来源：用户点击事件表event  更新策略：实时更新 | | |

## 用户推荐结果

|  |  |  |  |
| --- | --- | --- | --- |
| **表名** | **user\_recommend** | **表类型** | **hbase** |
| **主键、列族** | **字段名称** | **字段解释** | **说明** |
| rowkey |  | 主键 | device\_id |
| timestamp |  | 时间戳 | 发布时间 + 3位毫秒数 |
| recmd |  | 列族 | VERSIONS => 300，TTL => 30天 |
| tip | 待推荐资讯 | 存储资讯ID，TTL => 3天 |
| his | 已推荐资讯 | 存储多个资讯ID，逗号分割 |
| 其他 | 更新策略：每小时更新  计算过程：见[推荐计算过程](#_推荐计算过程) | | |

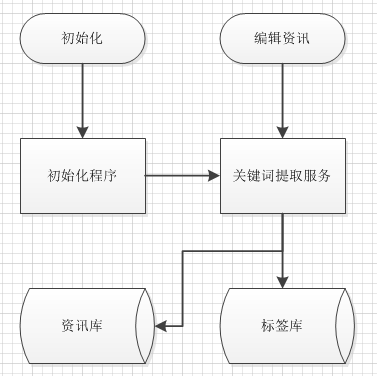
redis存储

推荐列表: list存储，key = “tip-“ + uuid

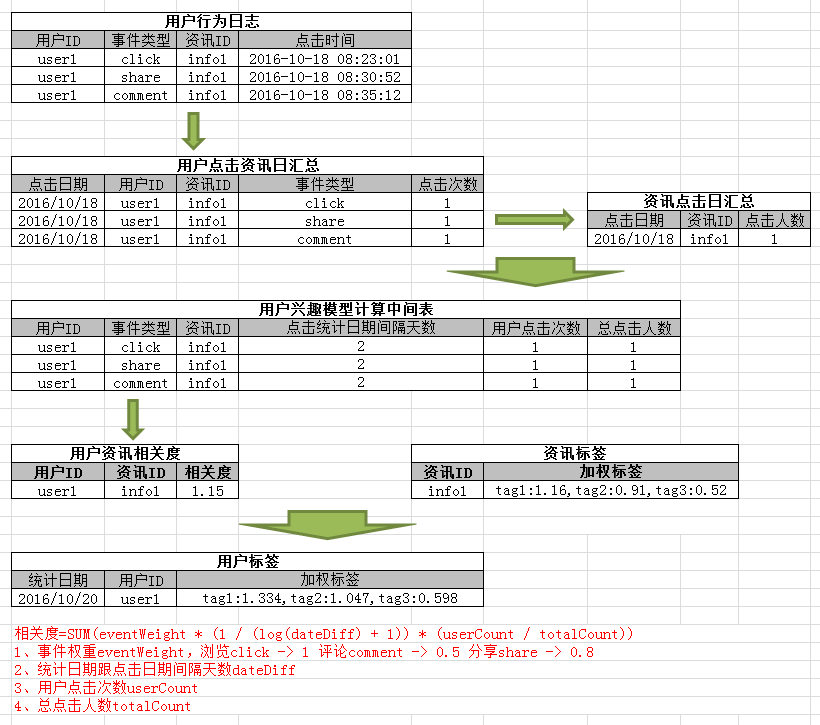
已推荐列表: set存储，key = “his-” + uuid

# 流程设计

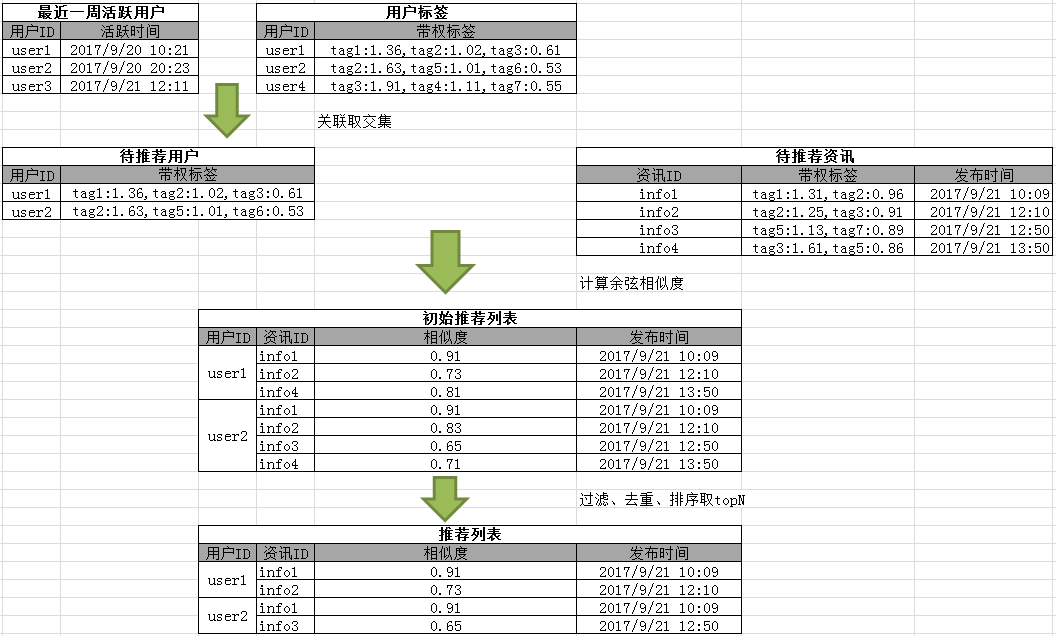
## 关键词提取流程



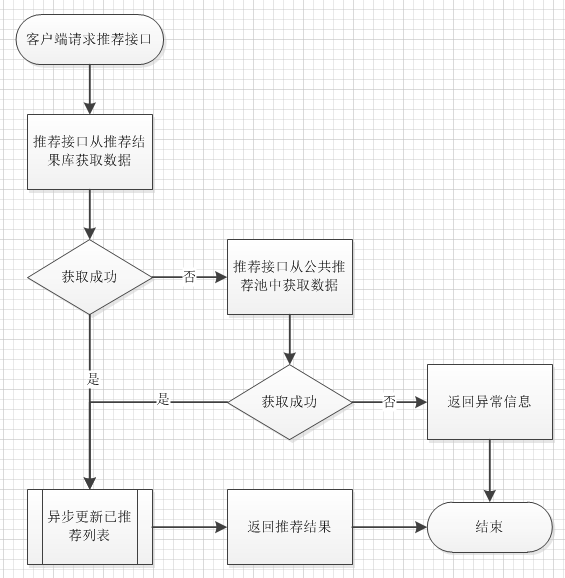
## 用户标签计算过程推导



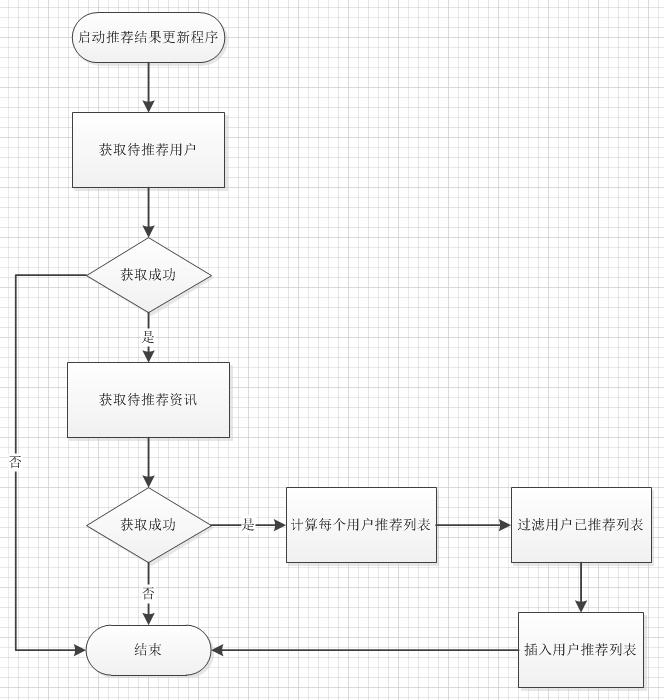
## 推荐计算过程



## 推荐接口调用流程



## 推荐结果更新流程



# 未完成

## 推荐消费接口

## 公共推荐池

当前实时热点资讯统计

# 下一步计划

## 推荐系统优化

算法优化，推荐流程优化，数据存储优化等

### 相关资料

[1] <http://www.rowkey.me/blog/2016/04/07/up-recommend/>