移动新媒体平台内容推荐方案系统设计文档

2017.10.11

**深圳市九指天下科技有限公司**

**目录**

[移动新媒体平台内容推荐方案系统设计文档 1](#_Toc496804177)

[1. 编写目的 3](#_Toc496804178)

[1.1 术语与缩略词 3](#_Toc496804179)

[1.2 对象及范围 4](#_Toc496804180)

[1.3 总体设计原则 4](#_Toc496804181)

[1.4 总体技术路线 5](#_Toc496804182)

[1.5 尚未解决的问题 6](#_Toc496804183)

[2. 推荐系统产品需求分析 6](#_Toc496804184)

[3. 推荐系统架构图 6](#_Toc496804185)

[3.1 业务架构图 7](#_Toc496804186)

[3.2 技术架构图 8](#_Toc496804187)

[4. 标签库 9](#_Toc496804188)

[4.1 核心概念 9](#_Toc496804189)

[4.2 标签提取流程 10](#_Toc496804190)

[4.3 特性 10](#_Toc496804191)

[5. 用户画像 10](#_Toc496804192)

[5.1 用户画像计算过程推导 11](#_Toc496804193)

[6. 生成推荐 12](#_Toc496804194)

[6.1 推荐计算过程 12](#_Toc496804195)

[6.2 名词解释 13](#_Toc496804196)

[6.3 关键技术 13](#_Toc496804197)

[7. 参考文献 13](#_Toc496804198)

# 编写目的

此概要设计说明文档的编写目的是为了说明内容推荐系统的体系架构，以及需求用例的各个功能点在架构中的体现，为项目组后续开展详细设计、系统建设与实现、系统测试提供依据。

本概要设计说明书适用于内容推荐系统详细设计、开发、测试等研究开发工作。

本文档的预期读者包括支持本项目的领导和负责人员；详细设计、开发、测试人员及文档编写人员；项目管理人员；参与项目实施的相关工作人员等。

## 术语与缩略词

|  |  |
| --- | --- |
| **词语** | **解释** |
| Hadoop | Hadoop is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models |
| HDFS | Hadoop Distributed File System |
| Hbase | HBase is the Hadoop database, a distributed, scalable, big data store |
| Spark | Spark is a fast and general engine for large-scale data processing |
| Hive | 基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供简单的sql查询功能 |
| Redis | 一个开源的使用ANSI C语言编写、支持网络、可基于内存亦可持久化的日志型、Key-Value数据库 |
| Scala | 一门多范式的编程语言，一种类似java的编程语言，设计初衷是实现可伸缩的语言、并集成面向对象编程和函数式编程的各种特性 |
| 余弦相似度 | 通过计算两个向量的夹角余弦值来评估他们的相似度 |
| 中文分词 | 将一个汉字序列切分成一个一个单独的词 |
| TextRank | TextRank 算法是一种用于文本的基于图的排序算法。其基本思想来源于谷歌的 PageRank算法, 通过把文本分割成若干组成单元(单词、句子)并建立图模型, 利用投票机制对文本中的重要成分进行排序, 仅利用单篇文档本身的信息即可实现关键词提取、文摘 |
| 业务架构 | 信息化总体架构的起点，业务架构定义了在总体企业战略指导下希望实现的业务能力及其关系。 |
| 应用架构 | 对信息系统进行的高层应用进行划分，指导整个信息系统构建和实施 |
| 数据架构 | 从跨组织应用系统的视角对数据进行组织和管理，包括对整个数据生命周期中数据的处理、存储、转换、整合、分布制定的策略、模型、流程、以及支持这些策略、模型、流程的技术架构方案。 |
| 技术架构 | 基于面向服务架构思路，实现信息、应用、流程横向贯通，支持应用架构优化和提升的技术平台架构；定义各个信息基础设施之间的关系，从宏观和微观角度分析信息系统的发展过程和技术要求，并为保障和支撑应用和数据提供一个可实现的基础。 |

* 1. 对象及范围

程序开发人员、系统架构师。

* 1. 总体设计原则
* 实用性原则

本项目的建设首先要坚持实用性原则，在实用的基础上考虑先进性和前瞻性，选用符合标准的，先进成熟的产品和开发平台，构建一个切合实际、解决实际问题的系统。

* 标准化原则

本项目具有数据量大、应用存有差异的特点，通过典型的应用功能与架构规划，才能为差异化的实施部署提供建设指导。

* 一体化原则

提供统一的功能与服务。

* 适用性原则

充分考虑不同现状，本系统应具备良好的可配置性和可扩展性，满足当前阶段行业现有的需求。

* 可维护性原则

在系统总体设计上注意系统的维护性，尽量采用成熟且易于维护的系统平台，维护应简单、易于操作。

* 安全性原则

系统建设应遵循互联网安全要求。

* 1. 总体技术路线
* 采用多层体系结构，严格按app和标准java开发规范进行开发。
* 提供灵活的界面展示和良好的扩展性。
* 系统界面简洁规范，美观大方，操作方便；具备人性化、简单的人机交互方式。
* 实现高级别的安全保障系统，充分保证数据和业务的安全性。

表格1 总体技术路线

|  |  |
| --- | --- |
| 分类 | 选型原则 |
| 技术选型 | * 前台界面展现技术采用Android、Objective-C； * 服务器端技术采用J2EE、JDBC、Servlet、MyBatis,Spring,Spark,Scala等； * Jdk版本1.7，Scala版本2.11.8，操作系统linux rhel5.7以上 |
| 开发平台 | * SG-UAP |
| Web服务器 | * Tomcat7.0以上版本、ngnix |
| 数据库 | * mysql5.5以上版本 |
| 开源软件 | * Spring、log4j、Hadoop,Kafka,Spark,Scala,Hbase,Hive,Redis等。 |

* 1. 尚未解决的问题

无

# 推荐系统产品需求分析

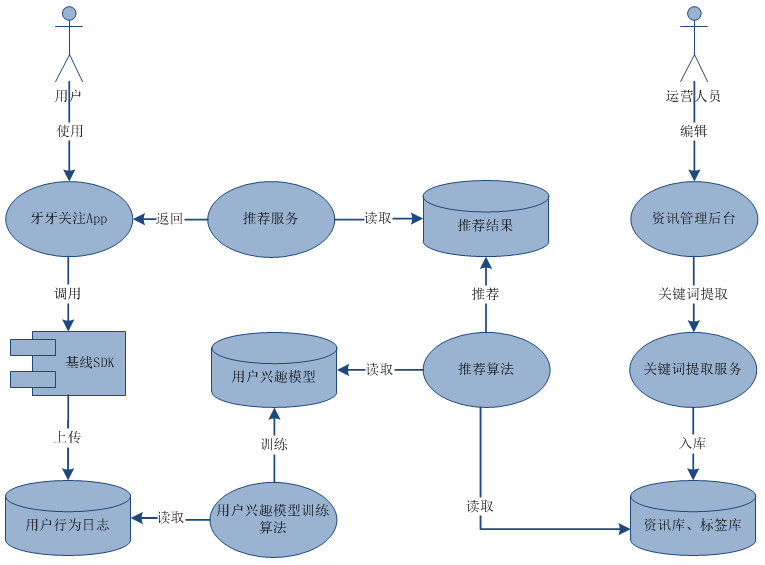
网络媒体中，新闻资讯的数量以近乎爆炸的速度增长， 使广大网络新闻的读者受到“信息过载”和“信息迷航”问题的困扰 。个性化新闻推荐系统可以通过分析用户使用习惯，识别用户兴趣点，将用户感兴趣的新闻资讯推荐给用户，过滤他们不感兴趣的内容，而推荐推荐系统就是为了有效地缓解上述问题。

基于内容的新闻推荐系统：它将计算得到并推荐给用户一些与该用户已选择过的项目相似的内容。例如，当你在网上看新闻时，你总是阅读与IT相关的文章，那么基于内容的新闻推荐系统就会给你推荐一些IT方面的文章。

# 推荐系统架构图

本发明基于MMSeg算法中文分词、TextRank算法关键词自动提取、高可用Hadoop分布式海量存储、高性能Spark分布式并行内存计算、分布式任务调度与监控告警等服务，打造一站式内容推荐方案，为用户提供即时、快捷、精准地个性化推荐服务。

## 业务架构图



**方案描述：**

**关键词自动提取**

通过采用TextRank关键词提取算法自动提取文本关键词，降低了人工成本，提高了关键词提取效率。

**分布式存储**

用Hadoop分布式文件系统作为底层存储系统，可以存储海量数据，扩容方便，数据多副本，安全性高，集群高可用。

**并行计算**

用户画像、推荐等算法用Spark并行计算框架实现，分析海量用户行为日志更加高效。

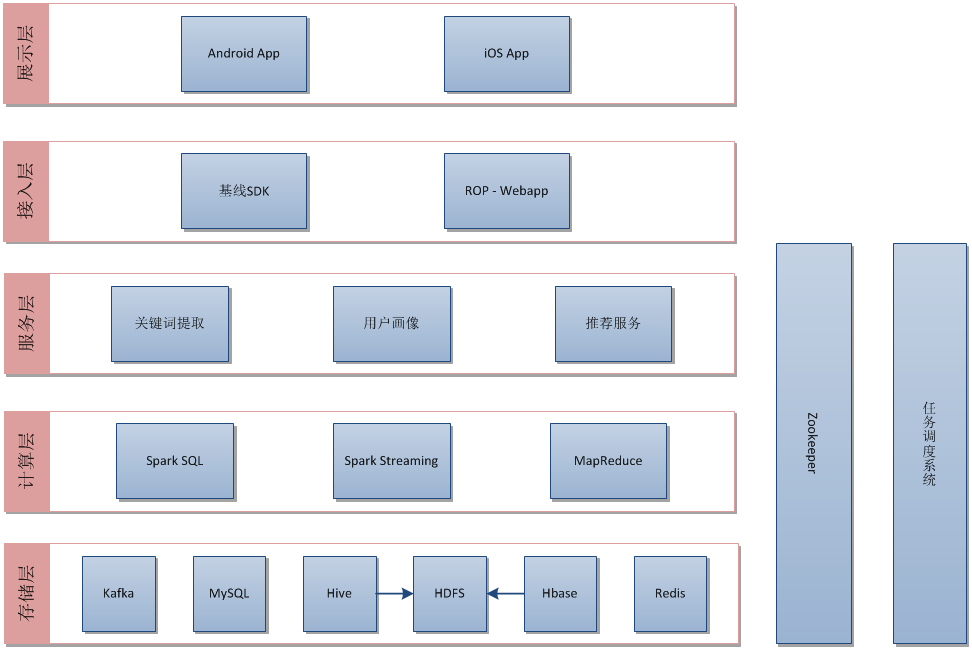
**Redis缓存**

推荐结果用redis缓存，推荐服务响应客户端请求更加高效。

**可扩展性**

存储系统、计算资源等都可以很方便的进行扩展。

## 技术架构图



**基线SDK**用于收集并上传用户点击行为日志，可以实时上传或者定时上传。

**关键词提取服务**提供中文分词、基于TextRank算法的关键词提取服务。

**用户画像服务**基于用户的历史点击行为数据，通过一定的算法计算得到用户兴趣模型。

**推荐服务**从推荐列表获取一定数量的资讯返回给客户端并记录已经推荐过的资讯。

**Spark**是并行计算通用框架，用户画像、推荐等算法都是基于该框架实现的。

**Kafka**用于临时存储基线SDK上传的用户行为日志。

**Hive**用于存储用户点击行为统计数据。

**Hbase**用于存储用户画像数据。

**HDFS**是Hive、Hbase表的存储系统。

**Zookeeper**用于协调Hadoop、Hbase、Spark、Kafka等服务。

**任务调度系统**用于调度用户行为统计任务、用户画像计算任务、推荐计算任务等。

**Redis**用于存储推荐结果数据。

# 标签库

首先对所有历史资讯（包含标题、摘要、内容）进行中文分词，过滤掉非名词、动词、形容词、副词等词性的词以及停用词，然后用TextRank算法提取资讯关键词，按相关度排名取topN作为资讯的标签并写入标签库。

## 核心概念

**标签（Tag）**

是联系用户与内容以及内容与内容之间的纽带，也是反映用户兴趣的重要数据源。是将其他实体转换为计算机可以理解的语言的关键的一步。

**标签库（TagLib）**

是对标签进行聚合的系统，包括对标签的管理、更新等。标签库的最终用途在于对用户进行行为、属性标记。

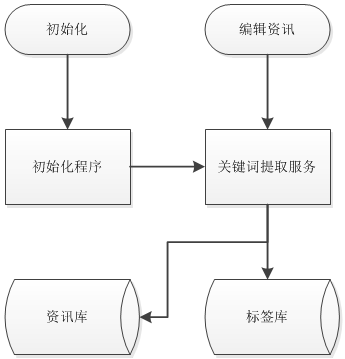
**中文分词（Chinese Word Segmentation）**

分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。我们知道，在英文的行文中，单词之间是以空格作为自然分界符的，而中文只是字、句和段能通过明显的分界符来简单划界，唯独词没有一个形式上的分界符，虽然英文也同样存在短语的划分问题，不过在词这一层上，中文比之英文要复杂的多、困难的多。

**TextRank**

TextRank 算法是一种用于文本的基于图的排序算法。其基本思想来源于谷歌的 PageRank算法, 通过把文本分割成若干组成单元(单词、句子)并建立图模型, 利用投票机制对文本中的重要成分进行排序，仅利用单篇文档本身的信息即可实现关键词提取、文摘。和 LDA、HMM 等模型不同, TextRank不需要事先对多篇文档进行学习训练, 因其简洁有效而得到广泛应用。

## 标签提取流程



## 特性

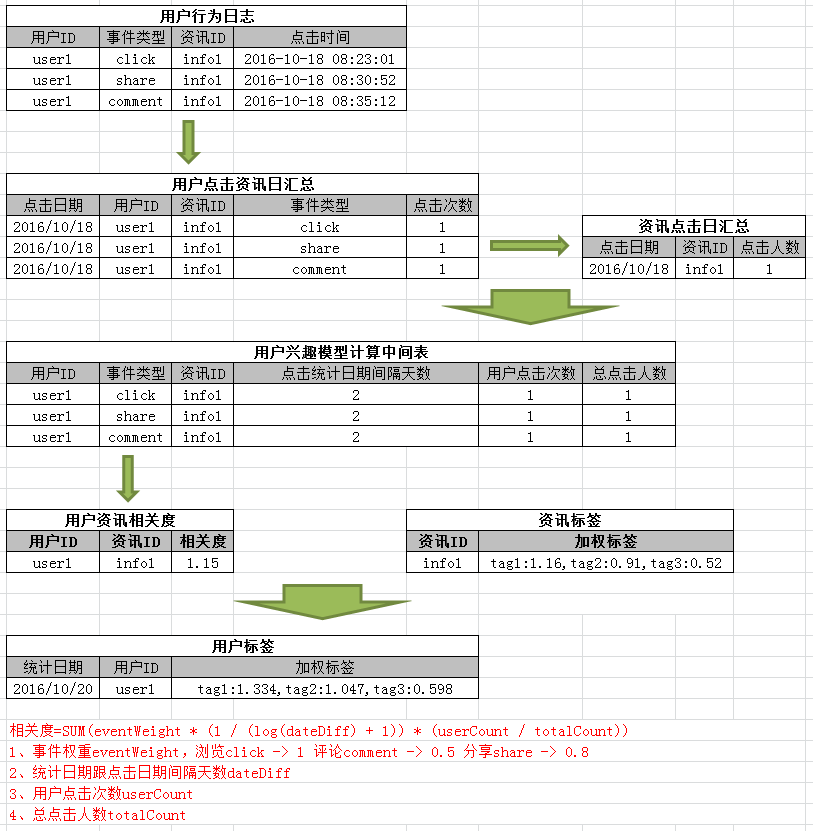
**关键词自动提取**

* **基于HMM中文分词**
* **关键词自动提取**
* **词库定时更新**
* **关键词提取服务效率高**
* **关键词提取个数可自定义**
* **标签、标签库可人工干预**

# 用户画像

用户画像即为用户打标签。通过用户的行为日志加上一定的模型算法计算得到用户的标签及每个标签的权重。

## 用户画像计算过程推导



**算法说明：**

1. 不同点击行为设置不同权重（浏览行为权重 < 评论行为权重 < 分享行为权重）。
2. 历史行为权重用点击日期与当前日期的日期间隔天数的自然对数做降权处理。
3. 对热点资讯做了降权处理。
4. 点击事件日志按天分区存储，并行读取，并行计算。
5. 资讯数据分区并行读取，并行计算。
6. 算法参数高度可配置，方便随时调整。

# 生成推荐

有了用户特征和内容特征，就可以关联用户和内容了，通过计算用户特征和内容特征的相似度得到推荐列表，这里采用了余弦相似度算法。

## 推荐计算过程



**算法说明：**

1. 推荐用户取最近一周的活跃用户，不活跃的用户也没有必要推荐，减少推荐计算量。
2. 提前过滤掉没有用户标签的用户，减少推荐计算量。
3. 采用了余弦相似度算法，算法抽象为一个静态方法，方便修改为其他相似度算法。
4. 生成最终推荐结果时会过滤掉已推荐资讯，降低推荐服务接口的复杂度。
5. 推荐结果分页后用redis存储，查询性能高。
6. 推荐结果设置TTL，过期自动清除，降低维护复杂度。
7. 算法参数高度可配置，方便随时调整。

## 名词解释

**余弦相似度**

余弦相似度，又称为余弦相似性，是通过计算两个向量的夹角余弦值来评估他们的相似度。

## 关键技术

**并行计算**

推荐算法用Spark Scala实现，用户数据、资讯数据等采用并行读取，并行计算，大大缩短了单次推荐任务计算时间。

**推荐结果用redis缓存**

**最终推荐结果分页后写入redis缓存，提高推荐服务效率，并设置TTL，过期自动清理。**

# 参考文献

1. <http://www.rowkey.me/blog/2016/04/07/up-recommend/>
2. <http://hadoop.apache.org/docs/r2.7.4/>
3. <http://spark.apache.org/docs/2.1.0/sql-programming-guide.html>
4. <http://www.redis.net.cn/tutorial/3501.html>
5. <http://nlp.csai.tsinghua.edu.cn/~lzy/publications/phd_thesis.pdf>