# NewsFeeds

## 项目概述

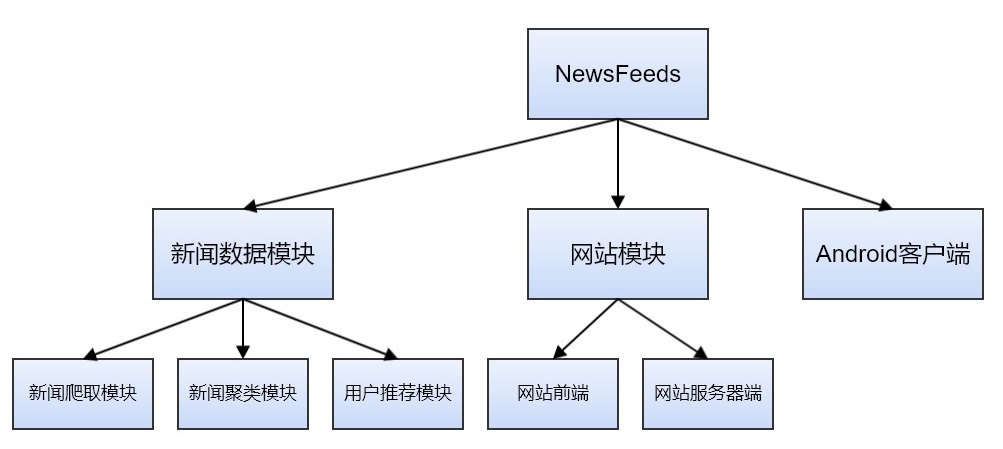
### 项目说明

本项目实现了一个新闻资讯收集、整理，进行个性化展示的网站。

网站服务器端使用Node.js作为服务器框架，使用MongoDB提供数据存储服务，并使用Redis作为数据缓存服务。在用户管理方面实现了邮箱认证注册登录，用户定制关键词等功能。

在新闻数据爬取上使用Python BeautifulSoup框架爬取并解析各大主流新闻网站的新闻页面，获取所需要的新闻数据；在新闻数据处理上通过维基百科提供的语料库，使用Word2Vec等自然语言处理相关的工具进行中文关键词建模，实现了按新闻关键词进行聚类，并根据用户感兴趣关键词进行新闻个性推荐；同时，使用了微软认知服务提供的推荐API，用于针对用户的特定行为进行建模推荐。

### 项目模块划分



项目模块划分

* 新闻数据模块
  + 新闻爬取模块：使用Python BeautifulSoup框架爬取并解析各大主流新闻网站的新闻页面，获取所需要的新闻数据存入数据库
  + 新闻聚类模块：使用中文维基语料库建立关键词模型，并按新闻关键词进行聚类
  + 用户推荐模块：使用微软认知服务-推荐系统提供的RESTful API进行用户行为统计并推荐新闻
* 网站模块
  + 网站前端
  + 网站服务器端
* Android客户端

### 主要源代码结构

NewsFeeds  
├── Graphs 实际运行效果截图, 系统架构图等  
├── [README.md](http://README.md) 开发文档  
├── Scripts 新闻数据相关脚本  
│ ├── config.ini 配置文件  
│ ├── [modelUpdater.py](http://modelUpdater.py) 推荐系统相关实现脚本  
│ ├── [newsCrawler.py](http://newsCrawler.py) 新闻爬取脚本  
│ ├── [service.py](http://service.py) 服务器后台常驻进程脚本  
└── Server 新闻网站服务器  
├── app.js Node.js入口脚本  
├── bin 服务器启动脚本  
├── config.json 网站配置文件（服务器地址，数据库源等）  
├── package.json Node.js依赖模块声明  
├── public 网站静态文件  
├── routes Node.js路由  
├── scripts 网站后端逻辑实现  
└── views 网站前端实现

192 directories, 65 files

### 运行环境

* 操作系统：Windows/Linux/OSX
* 服务器软件：Node.js 6.9.0及以上
* 逻辑数据库：MongoDB 2.4.1及以上
* 缓存数据库：Redis 3.2.9及以上

### 使用说明

1.进入服务器所在目录 Server

2.根据各个服务器配置，按如下说明更改配置文件config.json：

{  
 "host": "服务器主机地址",  
 "port": "服务器主机端口",  
 "connect": "MongoDB连接字符串",  
 "newscol": "MongoDB新闻数据库名",  
 "usercol": "MongoDB用户数据库名",  
 "mailoptions": {  
 "service": "邮件服务",  
 "email": "邮箱账号",  
 "password": "邮箱密码"  
 },  
 "redis": {  
 "port": "Redis端口",  
 "host": "Redis服务器地址"  
 },  
 "database": {  
 "host": "MongoDB数据库服务器地址",  
 "port": "MongoDB服务器端口",  
 "user": "Mongo用户名",  
 "password": "密码"  
 },  
 "api": {  
 "modelid": "微软认知服务-推荐系统模型ID",  
 "endpoint": "https://westus.api.cognitive.microsoft.com/recommendations/v4.0",  
 "token": "微软认知服务认证Token"  
 },  
 "genres": {  
 "personal": "推荐",  
 "hot": "热门",  
 "society": "社会",  
 "domestic": "国内",  
 "global": "国际",  
 "technology": "科技",  
 "finance": "经济",  
 "war": "军事",  
 "education": "教育",  
 "car": "汽车",  
 "game": "游戏",  
 "discover": "探索",  
 "entertain": "娱乐",  
 "fashion": "时尚",  
 "health": "健康",  
 "history": "历史",  
 "mobile": "数码",  
 "sport": "体育",  
 "travel": "旅游"  
 }  
}

3.安装Node.js依赖包：

npm install

4.启动服务器

npm start

## 新闻数据模块

### 新闻数据库设计

由于本项目使用基于文档存储的MongDB，每一个新闻使用一种类似JSON的数据类型存储，不需要有固定的结构（S）。但为了便于后续的建模与显示，在新闻爬取模块需要将每个新闻网站爬取下来的数据按照以下字段及其对应含义存储到数据库中：

{  
 "\_id" : 新闻id,  
 "title" : 新闻标题（唯一索引项）,  
 "source" : 新闻来源,  
 "time" : 发表时间,  
 "abstract" : 新闻摘要,  
 "comments\_count" : 新闻评论数,  
 "favorite\_count" : 新闻收藏数,  
 "genre" : 新闻类别,  
 "has\_image" : 是否有图片,  
 "imgurls" : 图片链接(数组),  
 "keywords" : 标题包含关键词(数组),   
 "related\_words" : 其他相关关键词(数组),  
 "uploaded" : 是否已上传到微软认知服务模型  
}

### 新闻爬取模块

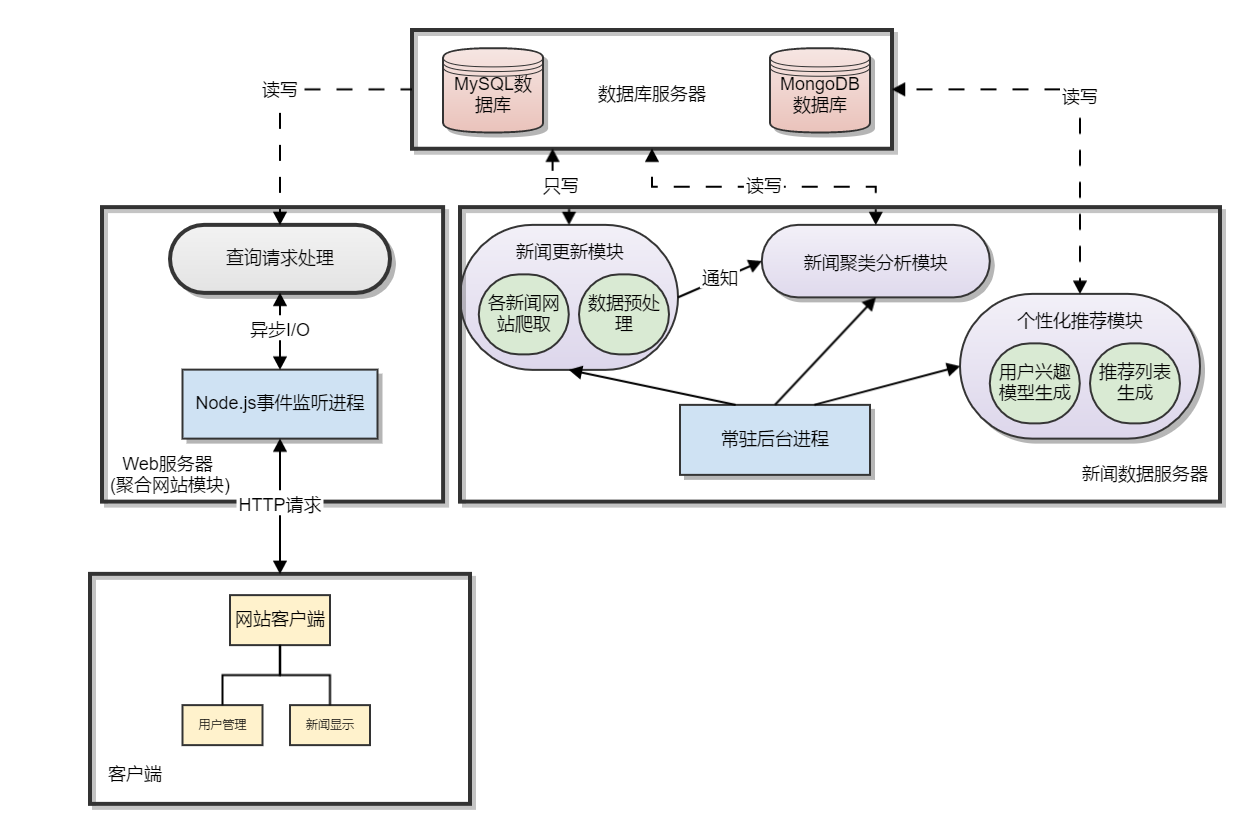
本项目中新闻爬取模块主要实现了对腾讯、网易、凤凰网、今日头条等四个网站的新闻数据爬取。

### 新闻聚类模块

### 个性化推荐模块

## 网站模块

### 网站架构



网站架构图

### 网站前端

本网站前端主要使用Bootstrap+jQuery框架实现。

### 网站后端

## Android客户端模块

## 成果展示