

软件需求分析规格说明书



新闻文本分类算法

团队名称： 卡其脱离太

指导教师： 代祖华

完成时间： 2021年5月18日

团队成员： 周学铭 何飞 常雅伦 谢林江

目录

[第一章 引言 1](#_Toc2700)

[1.1 编写目的 1](#_Toc31892)

[1.2 背景 1](#_Toc30099)

[1.3 参考资料 2](#_Toc17344)

[第二章 任务概述 2](#_Toc31336)

[2.1 目标 2](#_Toc19671)

[2.2 用户特点 3](#_Toc9350)

[2.3 项目拟采取的技术路线 3](#_Toc23031)

[第三章 需求规定 5](#_Toc18396)

[3.1 对功能的规定 5](#_Toc18545)

[3.1.1 系统整体功能模块 5](#_Toc18703)

[3.1.2 单条数据导入功能 5](#_Toc23696)

[3.1.3 多条数据导入功能 6](#_Toc19628)

[3.1.4 数据分析功能 6](#_Toc31852)

[3.1.5 操作记录功能 7](#_Toc11557)

[3.1.6 数据流图 7](#_Toc23167)

[3.1.7 ER图 8](#_Toc21907)

[3.1.8 软件系统状态图 8](#_Toc21985)

[3.2 对性能的规定 9](#_Toc20360)

[3.2.1 时间特性要求 9](#_Toc17742)

[3.2.2 灵活性要求 9](#_Toc1148)

[3.3 数据管理能力的要求 9](#_Toc19296)

[3.4 其他专门需求 10](#_Toc30618)

[第四章 运行环境的规定 10](#_Toc1450)

[4.2 软件支持 10](#_Toc8949)

[第五章 项目WBS图 11](#_Toc13714)

# 引言

## 编写目的

编写新闻文本分类算法系统需求规格说明书其目的是为用户需求与需求分析人员，软件开发人员以及相关利益设计人员的需求达成共识，以此保证软件开发的质量和需求的完整性。

新闻文本分类算法系统实现通过卷积神经网络对输入的新闻文本进行较高准确率的分类。能够实现用户通过单条新闻输入，或者批量新闻输入，然后通过卷积神经网络算法输出分类结果，准确率达到80%以上。在分类结果出来后，可以将分类结果通过可视化界面显示出来。最后需要一个报表功能记录之前的各种操作，用来保存记录。

## 背景

1. 项目名称: 新闻文本分类算法
2. 项目产品名称：新闻文本分类算法平台
3. 项目的任务开发者：周学铭，何飞，常雅伦，谢林江；
4. 项目的用户：新闻读者，新闻相关的工作者，学者，算法爱好者；
5. 项目与其他软件或其他系统的关系：数据来源于各大新闻媒体，该项目为各大新闻媒体提供新闻分类功能，并进行推荐。
6. 社会需求：新闻发展越来越快，每天各种各样的新闻令人目不暇接，对新闻进行科学的分类既能够方便不同的阅读群体根据需求快速选取自身感兴趣的新闻，也能够有效满足对海量的新闻素材提供科学的检索需求。
7. 技术可行性分析：Java目前已存在相当成熟的前后端框架，开发可视化界面较为方便。Python爬取数据的能力卓越。目前工业界上主流的分类算法还是较为传统的数据挖掘算法。而在学术界将CNN运用于文本分类的技术已较为成熟。
8. 法律可行性分析：Java的框架大多是开源的，没有知识产权相关条件的约束。

## 参考资料

1. Kim Y . Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.
2. 杜思佳, 于海宁, 张宏莉. 基于深度学习的文本分类研究进展[J]. 网络与信息安全学报, 2020(4):1-13.
3. 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017(6).
4. 于游, 付钰, 吴晓平. 中文文本分类方法综述[J]. 网络与信息安全学报, 2019(5).
5. 张海藩,吕云翔. 软件工程[M].人民邮电出版社:, 201309.348.
6. 张海藩,吕云翔. 实用软件工程[M].人民邮电出版社:, 201505.342.

# 任务概述

## 2.1 目标

本项目致力于设计一种能够输出分类的准确率不低于80%，且可以提供简单的可视化界面。并能完成以下功能：

1. 能够输入单条新闻。
2. 批量输入新闻。
3. 支持本地上传csv/xlsx文件。
4. 输出新闻分类结果。
5. 可视化显示分类结果。
6. 报表显示。

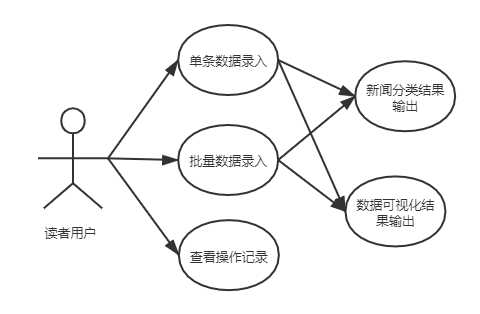


图2-1 用例图

## 2.2 用户特点

a. 读者用户希望能够快捷的显示的新闻分类，用来更好的检索新闻,节省寻找新闻花费的成本。

b. 新闻门户用户希望能将网站内海量的新闻数据迅速并精准的进行分类，方便对读者用户进行分类推荐或者检索，节省读者的时间成本。

通过需求调研显示有超过80%的用户群体，急需一款这样的产品。

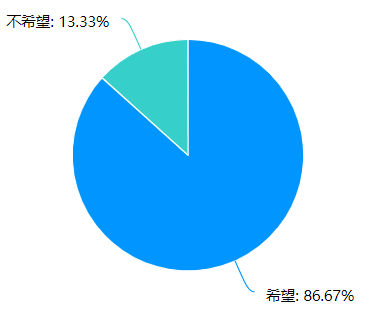


图2-2 用户对该产品的希望意愿

## 2.3 项目拟采取的技术路线

(1) 新闻文本分类过程

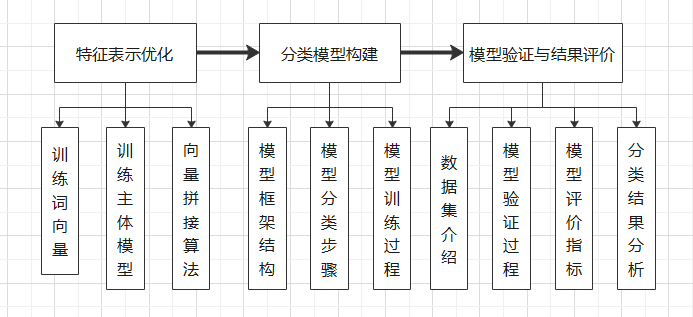


图2-3 新闻文本分类过程

1. 训练词向量：词向量的可以通过训练学习到词的分布式表示，给定n篇文章，就可以出词向量。关于词向量原理，先从神经网络语言模型(Neural Network Language mNNLM）谈起，NNLM直接通过一个神经网络结构对n元条件概率进行评估,元语言模型进行建模，通过词的n-1个历史词，来预测当前词。
2. 模型结构设计：基于CNN的新闻文本分类模型主要由输入层、卷积层、池全连接层和dropout分类器四层组成,模型结构图如图所示;

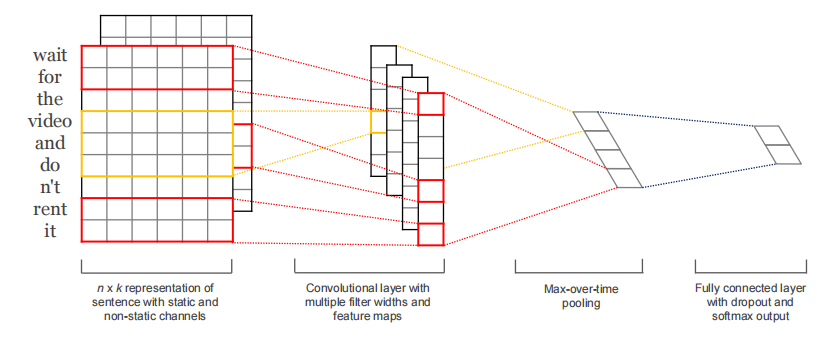


图2-4 卷积神经网络模型

1. 数据集：使用的新闻文本语料集为THU公开新闻数据集，数据内容和类样，适合于本文的文本分类任务。本文使用其中的10大类别，每个类别选择样本数量为8000条。
2. 系统设计流程

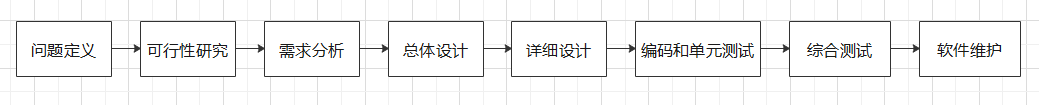


图2-5 系统设计流程

# 需求规定

## 3.1 对功能的规定

### 3.1.1 系统整体功能模块

新闻文本分类算法软件系统主要功能分为四个模块：分别是单条数据导入，多条数据导入，数据可视化分析，操作记录。

对于各个模块又对应不同的功能。

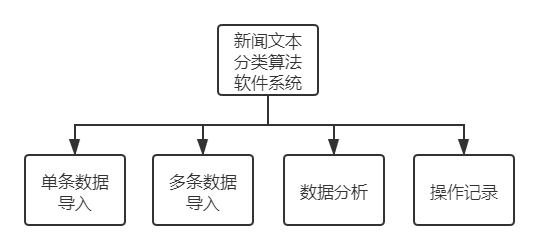


图3-1 总体设计

### 3.1.2 单条数据导入功能

单条数据导入功能模块主要是上传一个文件，并显示这个文件的信息。该功能实现了文件上传至数据库后台并存入数据库，实现了数据的永久性保存的功能，保存的文件信息直接存入数据库中，当再次需要这些内容时，直接从数据库中读出来即可。

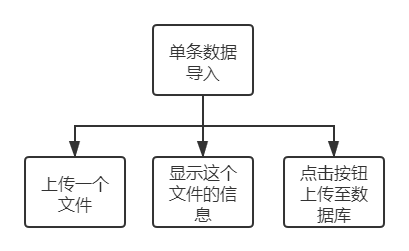


图3-2 单条数据导入功能

### 3.1.3 多条数据导入功能

批量数据导入功能模块主要是上传一批文件，并显示文件的信息。该功能实现了将一批数据以异步的方式直接上传至网站的后台，网站后台通过一步的方式，提取文件的内容，将文件的内容以数据流的方式存入数据库中，数据库永久性保存数据，直到前端显示数据或者后台需要处理数据时，才将数据加载，并进行指定的操作。

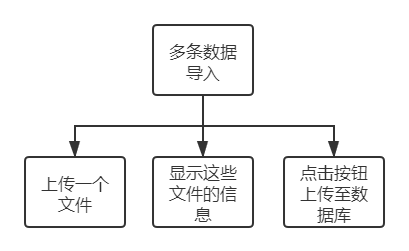


图3-3 批量数据导入功能

### 3.1.4 数据分析功能

数据分析功能模块主要是将数据库中的文件加载至前端界面，通过前端选择解析文件，选择好后，将选择的文件名传至后台，后台获取要解析的文件，开始采用文本分类算法解析文件，将文件进行解析，解析文件完成后，将文件解析后得到的内容传至前端进行显示。同时显示这些文件的真正类别，通过分析得到的结果和真正的结果进行对比，计算通过文本分类算法获得的数据的准确率，同时，显示在前端。

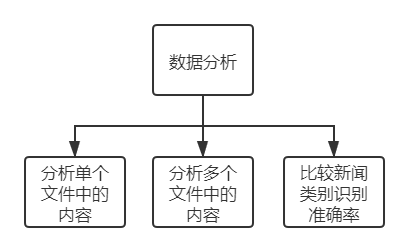


图3-4 数据分析功能

### 3.1.5 操作记录功能

我们的每一次访问都是一次操作，我们可能对数据库进行了写操作，可能上传了某些文件，可能从数据库中获取了某些文件，因此，时间的显示就显得格外的重要。我们的操作记录的模块主要就是负责这些内容的，我们不进行其它的操作，我们只记录用户进行了哪些操作，因此，这一部分可以直接独立于整个项目。

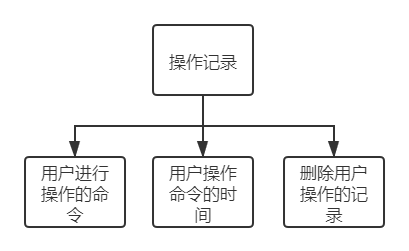


图3-5 操作记录功能

### 3.1.6 数据流图

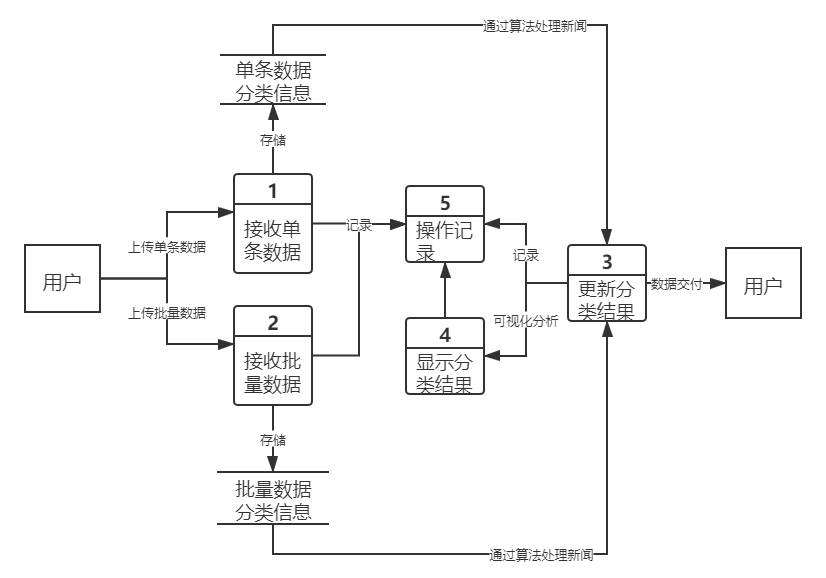


图3-6 数据流图

### 3.1.7 数据字典

表3-1 普通文件表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 表名：普通文件表(normal\_file) | | | | |
| 数据名 | 类型 | 范围 | 约束 | 备注 |
| file\_id | Integer |  | 主键 | 文件编号 |
| file\_name | String | 0-29个汉字 |  | 文件名 |
| file\_size | Integer |  |  | 文件大小 |
| file\_address | String | 0-29个汉字 |  | 文件地址 |

表3-2 新闻文件表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 表名：新闻文件表(news\_file) | | | | |
| 数据名 | 类型 | 范围 | 约束 | 备注 |
| news\_id | Integer |  | 主键 | 新闻编号 |
| news\_title | String | 0-29个汉字 |  | 新闻标题 |
| news\_content | String |  |  | 新闻正文 |
| news\_category\_right | String | 0-29个汉字 |  | 新闻正确类别 |

表3-3 分析结果表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 表名：分析结果表(analysis\_result) | | | | |
| 数据名 | 类型 | 范围 | 约束 | 备注 |
| news\_id | Integer |  | 主键 | 新闻编号 |
| news\_category\_analysis | String | 0-29个汉字 |  | 新闻类别（分析） |
| news\_category\_right | String | 0-29个汉字 |  | 新闻类别（正确） |

表3-4 统计分析表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 表名：统计分析表(statistic\_analysis) | | | | |
| 数据名 | 类型 | 范围 | 约束 | 备注 |
| news\_id | Integer |  | 主键 | 新闻编号 |
| accuracy | Integer |  |  | 正确率 |
| analysis\_time | Date |  |  | 分析时间 |

表3-5 操作记录表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 表名：操作记录表(operation\_record) | | | | |
| 数据名 | 类型 | 范围 | 约束 | 备注 |
| operation\_id | Integer |  | 主键 | 操作编号 |
| operation\_time | Date |  |  | 操作时间 |
| operation\_action | String | 0-29个汉字 |  | 操作动作 |

### 3.1.8 ER图

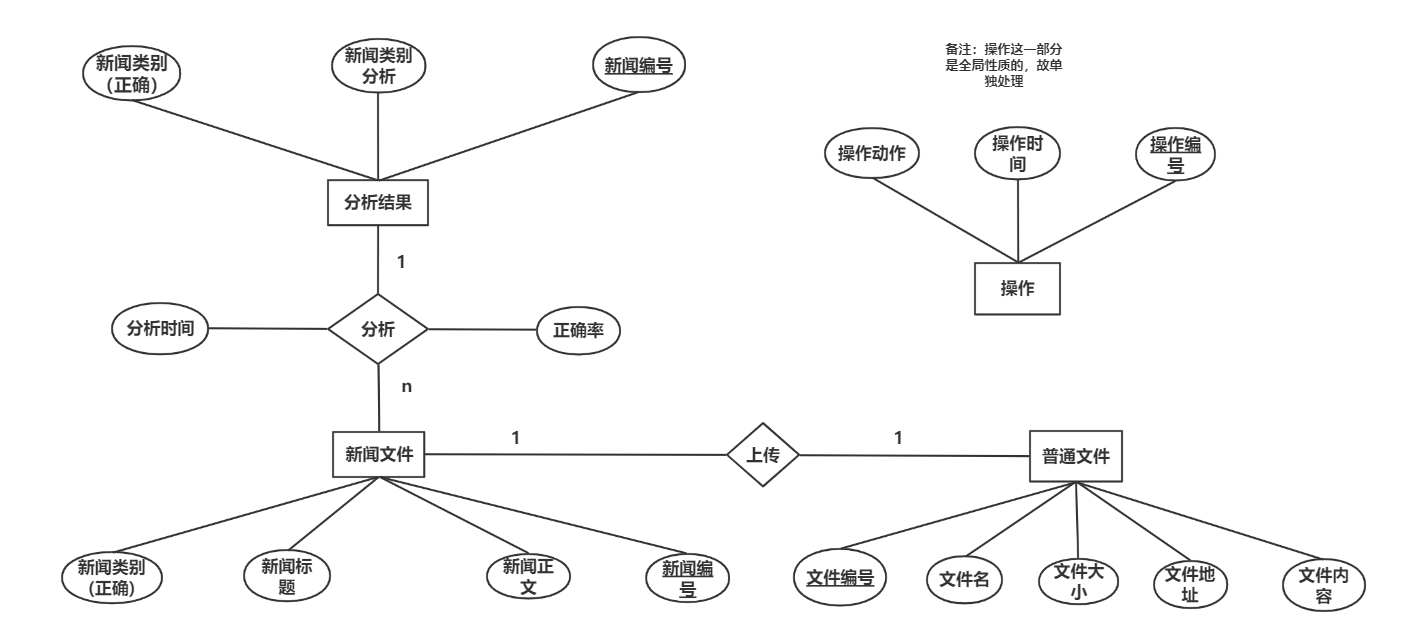


图3-7 ER图

### 3.1.9 软件系统状态图

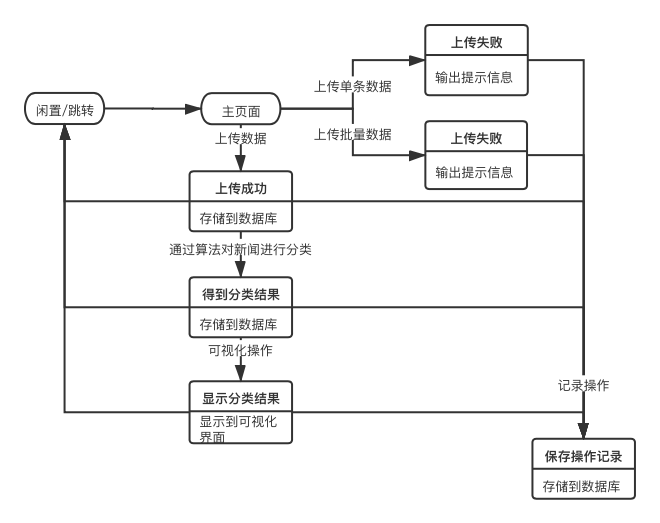


图3-8 软件系统状态图

## 3.2 对性能的规定

### 3.2.1 时间特性要求

（1）单条新闻，程序从输入到输出的执行时间不超过5s。

（2）可视化界面显示执行时间不超过2s。

### 3.2.2 灵活性要求

（1）操作方式的变化：用户根据人机界面提示进行简单的操作

（2）运行环境的变化：运行环境更改后，需要对网站重新配置，进行严格调试，整理运行环境配置文档

（3）灵活性和有效时限的变化：变化后需要重新修改用户使用文档

（4）计划的变化或改进：根据用户的需求不断的对软件进行升级和更新

## 3.3 数据管理能力的要求

本系统需要预先读取训练集的数据，并通过word2vec提取词向量，然后再通过设计好的CNN模型进行训练。

词训练模型为300x1\_word\_vecs\_word2vec，大小约为3.3G，所以要求服务器内存至少大于3.3GB。

训练集主要有七个，分别是MR,SST1,SST2,Subj,TREC,CR,MPQA。

表3-1 数据集

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data | Classes | Average sentence length | Dataset size | Vocab size | Number of words present in word2vec | Test size |
| MR | 2 | 20 | 10662 | 18765 | 16448 | CV |
| SST1 | 5 | 18 | 11855 | 17836 | 16262 | 2210 |
| SST2 | 2 | 19 | 9613 | 16185 | 14838 | 1821 |
| Subj | 2 | 23 | 10000 | 21323 | 17913 | CV |
| TREC | 6 | 10 | 5952 | 9592 | 9125 | 500 |
| CR | 2 | 19 | 3775 | 5340 | 5046 | CV |
| MPQA | 2 | 3 | 10606 | 6246 | 6083 | CV |

## 3.4 其他专门需求

（1）界面要求

界面的原则要求：方便、简洁、美观、一致。整个界面要保持友好、简易的风格。

输入设备：鼠标、键盘。

输出设备：显示器。

显示风格：IE、界面。

（2）安全性要求：对用户个人信息严格保护，对关键信息（操作记录等）进行加密处理和严格的权限进行管理限制。

（3）可维护性要求：本系统提供相关详细的开发和说明文档，需要更改或者有错误时,可以提供及时更改修复工作。

（4）易读性要求：界面符合简单大方风格，能合理显示用户的操作选项和正确结果要求，提示用户下一步的操作步骤，还有包含用户帮助文档。

# 运行环境的规定

4.1 设备

CPU:无要求；

显卡：要求为独立显卡(训练模型需要GPU)；

内存:4GB及以上;

硬盘:硬盘10G以上。

## 4.2 软件支持

模型训练：Anaconda 3；Python 3.5；Trensorflow 2.0。

数据库:mysql 5.5以上。

集成IDE: IntelliJ IDEA。

JDK版本:jdk1.8。

操作系统: Windows 10。

浏览器：谷歌浏览器。

# 项目WBS图

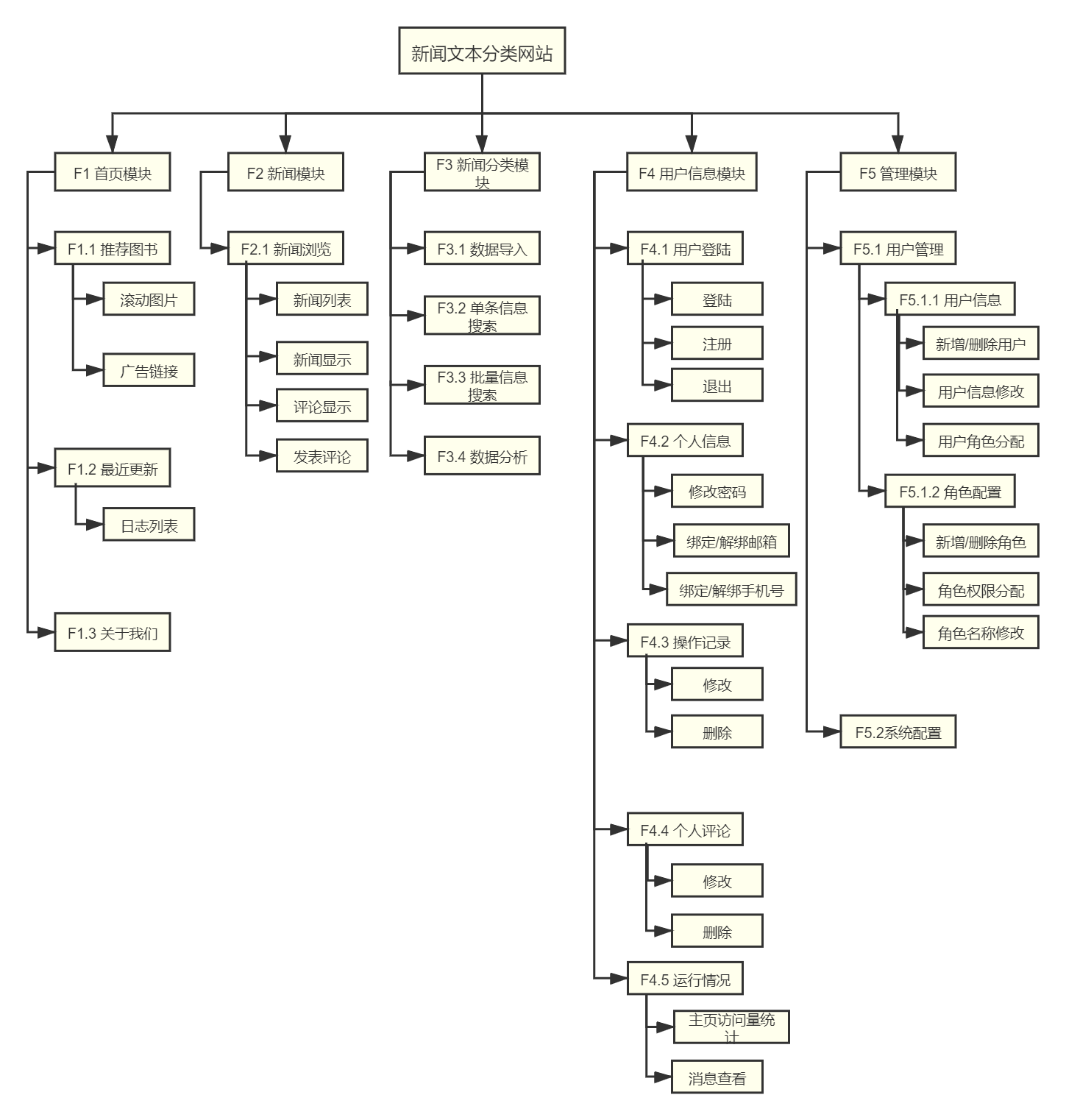


图5-1 项目WBS图