

# 基于微博的情绪可视化系统

## 项目详细方案

银杏大道东小组

2014/08

## 目录

1.	引言.....	4
1.1	编写目的 .....	4
1.2	背景.....	4
1.3	术语定义 .....	4
1.4	参考资料 .....	5
2.	需求分析.....	6
2.1	功能性需求.....	6
2.2	非功能性需求.....	8
2.3	可行性分析.....	9
2.3.1	核心技术实现可行性.....	9
2.3.2	资源可行性 .....	9
2.3.3	法律可行性 .....	9
2.3.4	操作可行性 .....	10
2.3.5	实用性和未来可拓展性.....	10
3.	概要设计.....	11
3.1	总体设计 .....	11
3.1.1	架构设计 .....	11
3.1.2	运行环境.....	13
3.1.3	基本设计概念和处理流程 .....	14
3.1.4	功能模块结构 .....	15
3.1.5	关键技术介绍 .....	18
3.1.6	主流程.....	21
3.1.7	系统实现方案 .....	22
3.1.8	功能需求与程序的关系.....	23
3.1.9	人工处理过程 .....	23
3.2	数据库设计.....	25
3.2.1	概述.....	25
3.2.2	逻辑结构设计 .....	26
3.2.3	物理结构设计 .....	27
3.2.4	数据结构与程序的关系.....	33

3.3	关键算法设计.....	33
3.3.1	垃圾微博过滤算法.....	33
3.3.2	情感分类算法 .....	34
3.3.3	微博影响力评估算法.....	36
3.3.4	关键字提取算法.....	36
4.	详细设计.....	38
4.1	包设计 .....	38
4.2	类设计 .....	39
4.3	模块设计 .....	43
4.3.1	情感分析模块 .....	43
4.3.2	噪声过滤模块 .....	44
4.3.3	可视化模块 .....	45
4.4	关键类说明.....	46
4.4.1	JunkKeywordBuilder 类说明.....	46
4.4.2	FeelingBuilder 类说明 .....	49
4.4.3	TopicfeelingBuilder 类说明.....	52
4.4.4	SentiAnalyzer 类说明.....	55
4.4.5	WeiboContentFilter 类说明.....	59
4.4.6	JunkWeiboFilter 类说明 .....	62
4.4.7	WeibomsgAction 类说明 .....	64
4.4.8	ViewScatter 类说明.....	66
5.	接口设计.....	69
5.1	接口设计概述.....	69
5.2	用户接口 .....	69
5.3	外部接口 .....	70
5.3.1	软件接口 .....	70
5.3.2	硬件接口 .....	70
5.4	内部接口 .....	70
6.	运行设计.....	71
6.1	运行模块组合.....	71
6.2	运行控制 .....	71

6.3	运行时间 .....	71
7.	系统出错处理设计.....	72
7.1	出错信息 .....	72
7.2	补救措施 .....	72
7.3	系统维护设计.....	72
8.	运行效果展示 .....	73
9.	附录.....	79
A:	心情-颜色对照表 .....	79
B:	情感分类表 .....	80
C:	团队组成与角色分工.....	80

# 1. 引言

## 1.1 编写目的

本项目解决方案说明书编写的目的是说明系统的整体架构以及程序模块的设计考虑，包括程序描述、输入/输出、算法和流程逻辑等，为软件编程和系统维护提供基础。本说明书的预期读者为系统设计人员、软件开发人员、软件测试人员和项目评审人员。

## 1.2 背景

在物质生活越来越丰富的今天，人们愈发关心自己的精神生活，每天在社交网络吐槽、po 图的我快乐吗？其他人的心情如何？能否通过一种有趣的方式来展示一些与用户情绪有关的信息和数据？起初，本团队只是想通过一种易于理解、且不拘泥于数字的形式，帮助用户了解他们所处的网络环境的一方面，即我们每天的感受如何，于是便诞生了本项目的雏形，一些运动的小点来显示不同心情的微博。进一步，我们发现这些心情或许有内在的规律，而这背后的价值是巨大的，于是在经过一系列前期数据比对后，发现用户的心情与时间、性别、地理位置这几个方面的关联最大，且有一定规律可寻，于是我们选取这三个维度对心情进行分析。

由上述需求，可以知道本项目的潜在价值是很大的，既可以为用户带来兼顾趣味与信息的数据，又可以为数据背后的潜在商业价值，提供准确可靠的分析挖掘。

## 1.3 术语定义

名称	解释
情感分类	情感的类别，一共为7大类，23小类。详见附件《情感分类表》
情感词	表达人的内心感受、情绪等的词称为情感词，例如“高兴”，“惊讶”等。
停用词	在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉某些字或词，这些字或词即被称为 Stop Words（停用词）。
否定词	表示否定意义的词，例如“无”，“没有”等。
程度副词	是对一个形容词或者副词在程度上加以限定或修饰的副词。一般位置在被修饰的形容词或者副词之前。例如“很”，“非常”，“相当”等。
情感模型	用于情感分类的基础情感单元，由（情感词，程度词，否定词）组成的三元情感特征模型。
情感词典	一个中文情感词汇资源。描述了一个中文词汇或者短语，包括词语情感类别、情感强度及极性等信息。
微博话题	新浪微博中用“##”标识的话题。

微博关键字(词)	同微博话题。
话题评论	包含“##”话题的微博中，用户的微博内容。
微博影响力	衡量一个微博账号每天在微博中影响力大小的数字，通过发微博情况、被评论、被转发的情况以及活跃粉丝的数量，综合评定一个账号的影响力大小。
垃圾微博	包含垃圾或无效信息的微博，例如广告、营销信息，第三方自动发布的推销微博。
垃圾词	从垃圾微博中提取出的关键词。如“拔草”，“淘宝”，“代理”等。
噪声词库	存放垃圾词的词库。
微博信息	一条微博文本内容。

## 1.4 参考资料

- [1] 廉捷，周欣，曹伟，等．新浪微博数据挖掘方案[J]．清华大学学报(自然科学版)，2011，51(10):1300-1305.
- [2] 朱嫣岚．文本情感倾向分析若干问题研究[D]．复旦大学，2007.
- [3] 朱力．中文词语情感倾向研究[D]．哈尔滨工业大学，2009.
- [4] 陆文星，王燕飞．中文文本情感分析研究综述[J]．计算机应用研究，2012，29(06):2014-2017.
- [5] 周延泉．基于情感词词典的中文句子情感倾向分析[D]．北京邮电大学，2011.
- [6] 朱杰，刘功申，陈卓．中文文本倾向性分类技术比较研究[J]．信息安全与通信保密，2010(4):56-58.
- [7] 魏韡，向阳，陈千．中文文本情感分析综述[J]．计算机应用，2011，31(12):3321-3323.

## 2. 需求分析

### 2.1 功能性需求

本系统通过微博 API 与网络爬虫结合的方式获取微博信息，并进行过滤处理后，利用自建的情感词典对微博信息进行情绪分类与统计，实现了微博用户的情感分类展示、情绪倾向分析、热门话题情感走势及分析结果的可视化图表显示，功能包含微博心情散点图，微博情感走向趋势，情感性别对比图，情感分布地图等。用户可通过简单的点击与系统交互，实现图像的响应与移动、设置观察条件来过滤信息等操作。

系统界面美观简洁、易于操作，同时，系统需最大限度地实现安全可靠、易于维护。

下表描述了本系统的主要功能性需求：

表 2-1 功能性需求表

功能模块	需求名称	详细说明
数据获取模块	获取微博信息	包括以下信息：消息 ID，用户 ID，用户名，屏幕名，用户头像，转发消息 ID，消息内容，消息 URL，来源，图片 URL，音频 URL，视频 URL，转发数，评论数，发布时间，赞数，地理坐标。
	获取微博话题信息	包括以下信息：消息 ID，用户 ID，用户名，屏幕名，转发消息 ID，消息内容，来源，图片 URL，赞数，转发数，评论数，发布地点，发布时间。
	获取微博用户信息	包括以下信息：用户 ID，屏幕名，性别，描述，地区编号，用户名，关注数，粉丝数，微博 URL，头像 URL
噪声过滤模块	获取垃圾微博	广告和营销微博。
	提取垃圾词	对垃圾微博进行分词处理，过滤停用词后提取关键词，统计词频后达到阈值地列为垃圾词。 供构造垃圾词过滤器使用。
	构造垃圾词过滤器	过滤掉垃圾微博。
情感分析模块	建立情感词典	将情感分类；建立词典并标注词语的强度、极性、情感分类
	建立停用词、程度副词、否定词、转折词词表	供文本分析使用
	建立情感模型	输入微博文本，输出文本的情感分类，情感极性值，强度。
	计算微博话题影响力	输入微博话题文本的赞数、评论数、转发数，输出该话题文本的影响力

文本处理模块	中文分词	将一个汉字序列切分成一个一个单独的词，并标注词性。
	构造信息过滤器	输入微博文本后，过滤掉表情符如“^^”，用户名如“@银杏大道”，转发“//”，网页链接“http://t.cn/RPQokAY”，话题“#八分青年#”
	提取关键词	输入微博文本，输出文本关键词
I/O 模块	文件读写	文件读写操作
	数据库读写	数据库读写操作
数据分析模块	按分类统计心情	按照性别、地理位置、时间，统计不同种类心情的用户数；不同情感极性（正面、负面、中性）的用户数。
可视化模块	情感分析流程展示图	显示了微博信息从收集、去噪、简单语义分析、情感分析到可视化的整个流程。
	心情/话题走势图	显示了一个话题在一个时间段内人们对它情感变化（褒贬）的走势。
	浏览微博心情散点图	点击导航 overall 按钮，后台程序从数据库中查询出最新的微博（700-800 条），以 json 的形式传回前台
	排序/打乱散点图	点击散点图节目 order 按钮 Isorder，变量变为 true/ false，每种心情的散点数量被计算，散点应该在的座标被计算，图像更新
	通过关键词查看相关微博	点击导航 time 按钮，后台程序从数据库中查询出每种心情 24 小时和一周的分布数量，以 json 的形式传回前台并展示
	浏览心情时间波动	点击导航 time 按钮，后台程序从数据库中查询出每种心情 24 小时和一周的分布数量，以 json 的形式传回前台并展示
	切换至心情/极性值时间波动	点击折线图上方的 change mode 按钮，后台，程序从数据库中查询出每种极性值/心情 24 小时和一周的分布数量，以 json 的形式传回前台并展示
	浏览心情性别对比	点击导航 gender 按钮，后台程序，从数据库中查询出每种心情在不同性别人群中的数量，以 json 的形式传回前台并展示
	浏览心情地理分布	点击导航 geo 按钮，后台程序从数据库中查询出每种心情在不同地区的分布数量，及对比颜色值，以 json 的形式传回前台并展示
	日/夜模式切换	点击切换按钮，调用 js 更改页面样式



## 2.2 非功能性需求

非功能性需求是指软件产品为满足用户业务需求而必须具有且除功能需求以外的特性，包括系统的性能、可靠性、可维护性、可扩充性和对技术和对业务的适应性等。

下表描述了本系统中的非功能性需求：

表 2-2 非功能性需求表

需求分类	需求名称	需求描述
用户界面需求	数据转化	将数据转换成图表、图形进行可视化展示、分析，界面友好，效果炫丽
	图标操作	支持各种图表操作：包括图表拖拽、导入/导出、放大/缩小、实体链接关系收缩/展开、关键实体强调等。
	多终端	支持在移动终端设备上数据进行可视化；
产品质量需求	性能	能够以 5 秒的最大响应时间,处理 50 个并发用户对本系统的访问，此时服务器的 CPU 占用率不超过 75%，内存使用率不超过 70%；系统网站的访问峰值时刻有 400 个用户，允许响应时间处长为 3 秒，此时服务器的 CPU 占用不超过 85%，内存使用率不超过 90%。
	可靠性	在给定的时间内以及规定的环境条件下，软件系统能完成所要求功能的概率。其定量指标通常用平均无故障时间和平均修复时间来衡量。
	安全性	主要涉及防止非法访问系统功能，防止数据丢失，防止病毒入侵和防止私人数据进入系统等。例如身份验证、用户权限、访问控制等都是与安全性相关的具体需求。
	易使用性	指本系统功能的简易程度，也包括对系统的输出结果易于理解的程度。
	可扩充性	指软件系统能方便和容易地增加新功能，通常用增加新功能时所需工作量的大小来衡量。
	可维护性	指在软件系统中发现并纠正一个故障或进行一次更改的简易程度。可维护性取决于理解、更改和测试软件的简易程度。
	互操作性	软件系统与其他系统交换数据和服务的难易程度。
	健壮性	指软件系统或是组成部分遇到非法输入数据以及在异常情况和非法操作下，软件系统能继续运行的程度。
	可移植性	指把一个软件系统从一种运行环境移植到另一个运行环境所花费的工作量的度量。

## 2.3 可行性分析

考虑到项目时间、资源等因素，在实际开发该系统时，常常要为资源不足和交付日期难以完成而苦恼，因而需要慎重地尽可能早的估计该课题的可行性。

可行性研究包括：经济可行性、技术可行性、法律可行性、操作可行性等。以下是对于本系统可行性的分析。

### 2.3.1 核心技术实现可行性

就目前使用的开发技术来说，系统的功能目标能够达到；利用现有的技术在规定的期限内开发工作基本能够完成。具体的技术分析参见下表：

表 2-3：参赛作品核心技术可行性分析

序号	核心技术关键	实现可行性
1	文本情感倾向性分析	人工构建情感词典，基于语义规则进行情绪分类。 采用中文分词技术及词性过滤，提取候选情感词及情感特征集合，通过情感词匹配和情感极性计算，获取文本的情感倾向性与心情分类
2	微博数据获取	通过新浪微博 API 与爬萌获取 JSON 格式源数据，解析 JSON，通过自建噪声库过滤垃圾数据，转化为本地数据模型后存入数据库
3	服务器端数据传输	采用 ajax 技术在服务器与浏览器之间进行数据交换，数据使用 JSON 进行封装与解析
4	前台数据处理	利用 Java Script 将后台传输过来的 JSON 数据解析后，转换为图形绘制所需要的数据
5	数据可视化	使用 HTML5 的 Canvas 标签特性，结合 JavaScript 库，将经过处理的数据绘制在前台页面上

### 2.3.2 资源可行性

- ♥ 时间：团队内所有成员都有精力投放在比赛中
- ♥ 人员：团队成员都具备充足的项目开发经验与较强的研究学习能力。
- ♥ 设备：使用的是主流笔记本，有服务器和客户端机器
- ♥ 开发地点：学校提供了实验室供团队开发使用

### 2.3.3 法律可行性

- ♥ 工具：本项目中使用的操作系统、开发工具等均为正版软件。

- ♥ 资料：技术资料来源于网络及相关书籍。
- ♥ 技术：本项目由本团队独立开发，使用的框架类库类型均为开源，技术上不存在侵犯专利等问题。

## 2.3.4操作可行性

本系统的用户界面为以 Web 网页的形式呈现，简单易用，只需用户完成点击操作和基本的文字输入，并在关键步骤会给出操作提示，用户只需了解浏览器的基本操作，即可轻松便捷的使用电脑或手机使用本系统。

系统管理员则需具备一定的计算机专业知识，经过培训后即可胜任。

## 2.3.5实用性和未来可拓展性

- ♥ 实用性
  - ◇ 基于活跃人数极多的社交平台微博开发，展现主题为大众喜闻乐见的“心情”，可以让用户轻松得到微博上的心情分布情况与发展规律，兼具趣味性与实用性。
  - ◇ 在 PC，智能手机，平板电脑等各种平台都可浏览本网站，实用性极强。
- ♥ 未来可拓展性
  - ◇ 提供了一种基于社交媒体的情感分析方法，可作为基础工具，为社会化媒体数据研究提供帮助。
  - ◇ 本项目中情感分析数据可作为商业决策的参考依据，分析方法可嵌入电影票房预测系统、商品推荐系统、用户个性化服务定制等，以提升结果的准确度。
  - ◇ 目前项目的数据来源是微博，将来可以扩大数据，研究不同社会化媒体用户的情绪状态。
  - ◇ 可以加入更多统计因素，获取心情与各种因素间的关联关系。
  - ◇ 可以进行个性化定制，获取专人的心情“晴雨表”
  - ◇ 可以根据情绪进行话题定位、舆情追踪等。

## 3. 概要设计

在本系统的需求分析阶段中，已经对本系统的功能需求做了详细的阐述。本阶段将会在需求分析阶段的基础上对本系统做进一步的概要设计，主要包括本系统的总体设计、操作处理流程设计、数据库结构的设计和关键算法设计等。以上系统模块的设计将结合需求分析阶段的功能需求，把各模块间的关系给建立起来，从而完成整个系统的概要设计需求。另外，在下一阶段的详细设计中，本阶段的概要设计将作为参考，以方便完成整个系统的设计工作。

### 3.1 总体设计

#### 3.1.1 架构设计

软件体系结构是构建计算机软件实践的基础。与建筑师设定建筑项目的设计原则和目标，作为绘图员画图的基础一样，一个软件架构师或者系统架构师陈述软件构架以作为满足不同客户需求的实际系统设计方案的基础。

下面将从几个方面介绍本系统的架构设计。

##### 3.1.1.1 整体架构

本系统分为展现层、业务逻辑层、中间层、数据库管理层。整体架构如图 3-1：



图 3-1 系统架构图

- **展现层**

展现层是该平台直接面向用户的界面，直接与用户互动，它的主要作用是提供给用户各种操作界面，各种操作方案，捕获用户的输入，通过表现层将这些用户输入的信息捕获并收集，以便交由到服务器进行处理，从而将其反馈给前台。用户将通过表现层的各个功能群组中的功能模块进行操作。前端表现层会将收到的请求以及各项数据收集起来，传递给应用层，进行相应的处理。

- **业务逻辑层**

业务逻辑层的主要任务是根据实际的业务规则，实现相应的业务逻辑功能。在这一层中，主要描述表现层所涉及到的目标服务进行对应的实现。在此业务逻辑层主要实现对应表现层的功能模块。

- **中间层**

本系统服务器中采用 Spring MVC 架构。Spring 架构提供了构建 Web 应用程序的全功能 MVC 模块。通过策略接口，Spring 架构是高度可配置的，而且包括多种视图技术，整个社交平台系统页面展示使用的 Java Server Pages (JSP) 技术。Spring MVC 分离了控制器、模型对象、分派

器以及处理程序对象的角色。Spring 的 Web MVC 架构是围绕 Dispatcher Servlet 设计的，它把请求分派给处理程序，同时带有可配置的处理程序映射、视图解析、本地语言、主题解析以及上载文件支持。Spring 提供了一个控制器层次结构，可以派生子类。Spring MVC 现在较成熟的将 web 页面中的输入元素封装为一个（请求）数据对象；根据请求的不同，调度相应的逻辑处理单元，并将(请求)数据对象作为参数传入；逻辑处理单元完成运算后，返回一个结果数据对象；将结果数据对象中的数据与预先设计的表现层相融合表现给用户。

● 数据层

数据库管理层主要职责是管理数据库之间的各种连接、断开等，并对数据库的各种操作中的一些异常进行记录和给出一些提示，以方便开发人员在需要的时候，对数据库相关的功能进行测试或者调试。一个系统中的所有功能，都涉及到各种各样的信息配置、业务处理数据、系统运行效果状态等，数据库对这些数据信息可以进行归档，提供一些基本的查询接口，以保证数据的可靠性和完整性。

3.1.1.2 部署架构

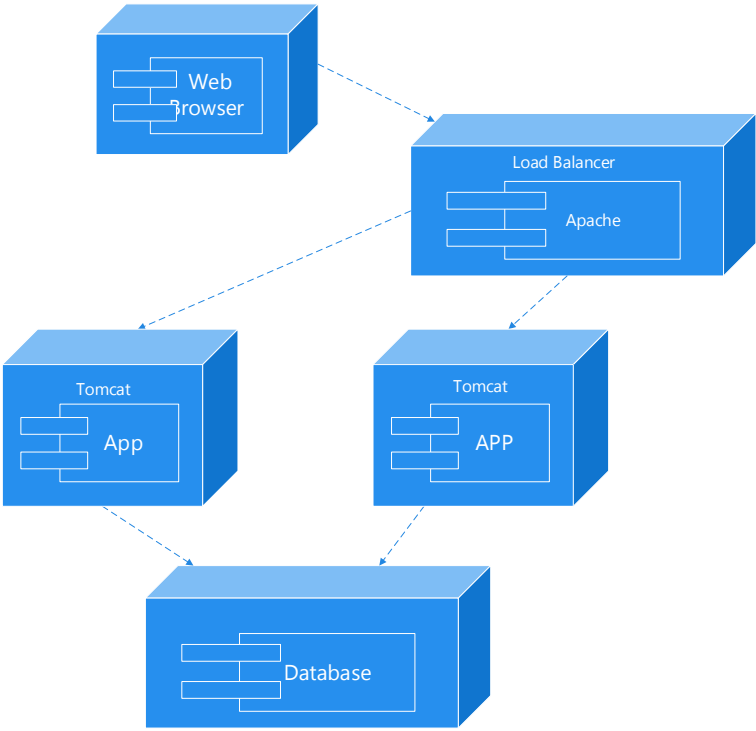


图 3-2 系统部署图

3.1.2 运行环境

3.1.2.1 软件环境

服务器端	
操作系统	Linux
数据库管理系统	MySQL

服务器	Tomcat 7.0
-----	------------

客户端	
主机	PC、IPad、Android、IOS 移动端
操作系统	Windows XP+/OS X/IOS 5+/Android
浏览器	IE 9+/Chrome/Firefox/Opera/Safari

### 3.1.2.2 硬件环境

服务器	最低配置
CPU	1.4GHz
内存	2G
硬盘	500G

### 3.1.3 基本设计概念和处理流程

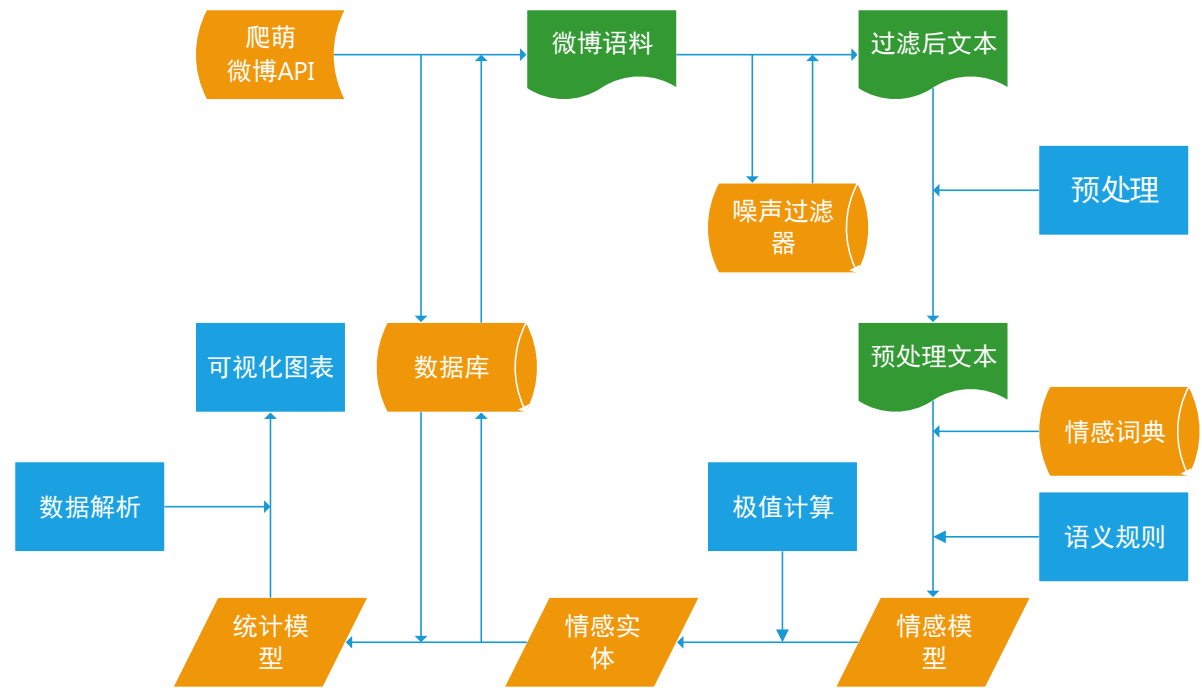


图 3-3 系统数据处理流程图

### 3.1.4 功能模块结构

功能模块是指数据说明、可执行语句等程序元素的集合，它是指单独命名的可通过名字来访问的过程、函数、子程序或宏调用。功能模块化是将程序划分成若干个功能模块，每个功能模块完成了一个子功能，再把这些功能模块总起来组成一个整体。以满足所要求的整个系统的功能。

#### 3.1.4.1 总体结构

根据前面的分析，本系统划分为两大部分：高层模块和底层模块。高层模块主要负责业务逻辑的实现，底层模块负责数据的读写和中文文本的处理分析，高层模块通过调用底层模块实现其功能，高层模块和底层模块之间的通信依赖于接口。

图 3-4 详细说明了本系统的模块划分结构：

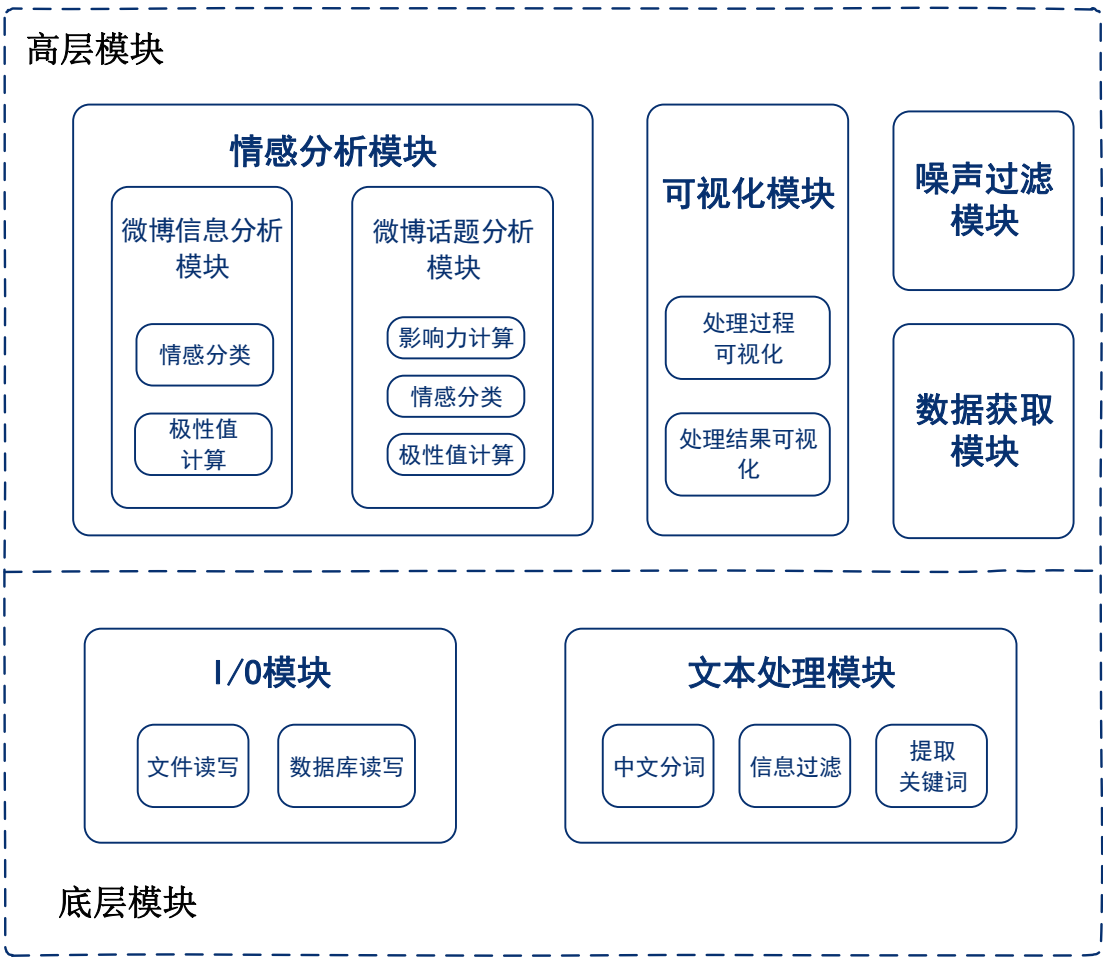


图 3-4 模块划分图

#### 3.1.4.2 情感分析模块

情感分析模块的作用是对微博内容进行情感挖掘。根据情感词典提取出微博文本中的情感词并根据语义规则提取出情感词的修饰词构筑相应的情感模型，根据情感模型计算情感极性值，对微博进行情感分类及情感倾向性判断。



根据信息主体的不同，本模块可分为两部分：微博信息分析与微博话题分析。微博信息分析指的是对用户发布的普通微博进行分析，微博话题分析的数据则来源于用户对一特定话题所发表的评论。在话题分析中，为更好的体现大众对话题事件的观感，在微博文本信息分析的基础上加入了微博影响力作为权重因数。

在本系统中，共将情感分为 7 大类 23 小类，部分分类示例如表 3-1 所示，完整分类请参照附录 B。情感倾向性则分为 3 类：正面、中性和负面。

表 3-1 情感分类示例表

编号	情感大类	情感类
1	乐	萌 (PM)
2	乐	快乐 (PA)
3	乐	喜爱 (PB)
4	好	尊敬 (PD)
5	好	赞扬 (PH)

### 3.1.4.3 噪声过滤模块

噪声过滤模块主要负责在微博内容识别阶段准确地识别出垃圾微博，并将垃圾微博从待情感分析的微博列表中删除。

噪声过滤需要先收集一定量的噪声微博作为训练文本，提取出关键词后统计频率，滤掉停用词后达到设定的阈值的关键词则列为垃圾词，录入噪声词库。

然后就是构造过滤器，本系统主要采取词语匹配的方式构造过滤器，通过“词共现”来识别出垃圾微博。

例如关键词只有“产品”时不能划为垃圾微博。

但关键词包括“微商#微信#产品#推广#”则为垃圾微博，例如：“淘宝靠信誉，**微商**靠人脉，不**推广**谁知道你**产品**好，不加入你的**产品**卖给谁。亲戚朋友的大米可不好赚，不管您现在生意好坏 只要是**微商**就需要不断注入新鲜客源。。这样才能长长久久。亲们排单继续……想要赶上微商旺季抓紧时间啦#琪柒专业**微信推广**团队 加**微信** Qqing1129#”

### 3.1.4.4 I/O 模块

本系统中很多地方需要与文件和数据库交互，进行读写操作。

I/O 模块封装了建立连接，获取对话，关闭对话等对数据库和文件的读写操作，提供给高层模块统一的读写接口。

### 3.1.4.5 文本处理模块

文本处理模块属于底层模块中的工具模块。主要向上提供了三个工具接口：中文分词、信息过滤和提取关键词。

中文分词负责将一个汉字序列切分成一个一个单独的词，并标注词语的词性。

信息过滤主要指过滤掉文本中的转发符“//”，用户名“@XXX”，微博话题符号“#XXX#”，网页

链接 <http://xxx.com.cn> 等。

提取关键词指提取出一个句子的关键词或关键短语。

3.1.4.6 可视化模块

可视化模块主要负责将微博情感挖掘的数据以图表的形式生动活泼的显示出来，使用户可以观察到最近一段时间，活跃在新浪微博上的用户的心情和新浪微博的热点话题及参与的评论。

对于整体业务功能的实现，主要包括处理过程可视化和处理结果可视化两个模块。图 3-5 具体描述了模块的划分和层次结构。

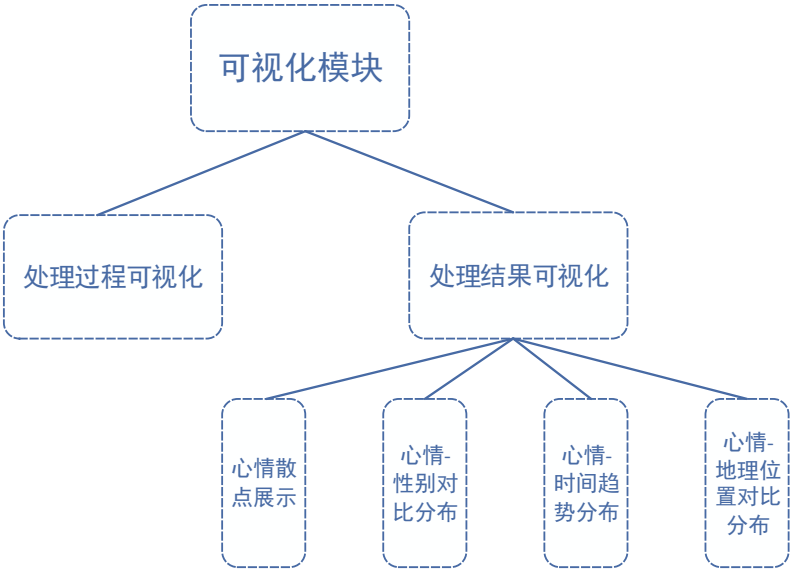


图 3-5 可视化模块业务图

处理结果可视化又分为四个模块，表 3-2 描述了各个模块的功能介绍。

表 3-2 可视化模块细化表

模块名称	功能介绍
心情散点图模块	展示包含发布者心情信息的微博内容散点图，点的颜色代表该条微博的情感类别，当用户的点击一个散点时，散点会运动至显示区域，显示该条微博内容，用户可以点击前往该微博作者的微博留言。
心情-时间模块	显示一周或每天内各心情随时间趋势变化的折线图。可按心情分类或情感极性显示。
心情-性别模块	显示男女所有心情分类的微博分布。通过饼状图和柱状图，用户可以清楚地看到不同性别人群中不同心情所占的比例。
心情-地理模块	系统将抓取到的微博内容和地理位置的信息相结合，显示在中国地图上。每个省级行政单位的颜色均为计算出的心情均值对应的颜色。

## 3.1.5 关键技术介绍

### 3.1.5.1 自然语言处理相关技术

因为微博文本情感分类的处理对象为中文文本，所以必然会跟自然语言处理的许多技术相关，主要包括分词和词性标注等。

#### 3.1.5.1.1 中文分词

分词就是将没有分割标记的字序列转换到符合语言实际的词串。词是自然语言理解中有意义的最小构成单位，但是中文文本词与词之间没有明显的界限标志，因此词语边界识别是中文文本处理首先要解决的问题。

现有的分词算法可分为三大类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。按照是否与词性标注过程相结合，又可以分为单纯分词方法和分词与标注相结合的一体化方法。

当前分词面临的问题有：一是歧义识别。同样的一个句子可能在多种切分方法，不同切分方法会得到不同的分词结果；二是未登录词的识别。未登录词是指未收录到分词词典但又确实表达一定含义、不能切分为更小的语义单元的词，如人名、地名、新词等。

本系统中采用的是目前较先进的 NLPIR 分词系统。

NLPIR 分词系统前身为 2000 年发布的 ICTCLAS 词法分析系统，从 2009 年开始，为了和以前工作进行大的区隔，并推广 NLPIR 自然语言处理与信息检索共享平台，调整命名为 NLPIR 分词系统。主要功能包括中文分词；词性标注；命名实体识别；用户词典功能；支持 GBK 编码、UTF8 编码、BIG5 编码。该系统的主要思想是通过 CHMM（层叠形马尔可夫模型）进行分词，通过分层，既增加了分词的准确性，又保证了分词的效率。

#### 3.1.5.1.2 词性标注

词性标注就是根据一个词在某个特定句子中的上下文，为这个词标注正确的词性。因为在中文中广泛存在着词语的词性多于一个的歧义现象。这种现象也叫做词性歧义或词性兼类。例如“好”这个词在“你好漂亮”和“他是个好人”中分别为副词词性和形容词词性。可见，词性标注的正确率对整个文本情感分类是非常重要的。目前词性标注的方法要有 3 种：基于规则的方法、基于统计的方法、统计与规则相结合的方法。

#### 3.1.5.1.3 情感词汇本体构建技术

我们以 DUTIR 情感词汇本体库及 HowNet 情感词典为基础，进行删改，并整合自己构建的热门网络情感词库而整理得到的一个中文情感词汇资源。该资源从不同角度描述一个中文词汇或者短语，包括词语词情感类别、情感强度及极性等信息。

情感词典的情感分类体系以 DUTIR 情感词汇本体库的情感分类体系作为参照，在该本体库 7 大类 21 小类的基础上根据现有网络环境添加两种小类。最终情感词典中的情感共分为 7 大类 23 小类。情感强度分为 1.0, 1.5, 2.0, 2.5, 3.0 五档，3.0 表示强度最大，1.0 为强度最小。

情感词典中，一般的格式如表 3-2 所示。

表 3-3 情感词典格式举例

词语	情感分类	情感分类序号	强度	极性
抓狂	NI	13	3.0	0
可爱	PM	1	2.0	1
自信	PH	5	2.0	1
害羞	NG	9	1.5	0

具体的情感分类如附录 B 所示。

构造该资源的宗旨是在情感计算领域，为网络环境下的中文文本情感分析和倾向性分析提供一个便捷可靠的辅助手段。

### 3.1.5.2 情感倾向性分析

情感倾向性分析，也叫做情感倾向性计算，根据其分析对象的不同，可以划分为词语级情感倾向性分析、句子或短语级情感倾向性分析、篇章级情感倾向性分析等。

#### 3.1.5.2.1 词语的情感倾向性分析

词语是构成文章的基础，对词语的倾向性进行分析是整个情感倾向性分析的前提，它的主要内容包：提取候选词，该候选词可能具有情感倾向；分析该候选词，判断其极性及其极性强度。一般来说，中文文本中具有情感倾向的词以名词、形容词、动词、副词为主。对词语进行情感倾向性计算主要有两种方法：基于词典的方法和基于语料库的方法。

#### 3.1.5.2.2 句子的情感倾向性分析

不同于针对单独的词语进行计算的词语级的倾向性计算，句子级情感倾向性分析主要是处理上下文中的语句。主要有模式匹配法、句子分类法、句法分析法。

◆ 模式匹配法

该方法主要是预先建立好一些已规定了倾向性及其极性强度的短语或句子模式；然后通过匹配这些句子，将符合规定好模式的句子挑选出来，赋予这些句子预先定义好的倾向性和强度。对于符合模式的句子，该方法具有较高的准确度，而对于无法匹配的句子和短语，则无法判断。

◆ 句子分类法

该方法是借鉴了文本分类的思想，预先定义好情感分类体系，将句子情感分析看成是一个分类的过程，然后通过分类器来进行分析。

◆ 句法分析法

这种方法结合了词语的情感倾向和句法的特点，首先进行句法分析，得出一个句子中的词语之间的关系，然后再结合词语的情感倾向，进而分析句子的情感倾向和强度。

3.1.5.2.3 篇章的情感倾向性分析

篇章级的情感倾向分析，即整体判断篇章的情感倾向性。其中篇章包括整篇文章或者部分段落。目前主要使用的方法有两种：文本情感分类和文本情感倾向计算方法。文本情感分类通过采用统计学方法，将文本类别定义为褒义、中性、贬义等。

文本分类需要有训练集，来对分类器进行训练，实际上就是构建一个文本分类器的过程。分类时首先通过收集一定数量的训练集：然后通过特征提取的方式，将训练文本表示成某种方式：接着使用训练文本对分类器进行训练，直到达到预定的条件为止：最后对于新来的文本，同样将其用特征提取的方式表示成某种方式，然后输入到分类器中，分类器的输出就是该文本所属的类别。可以看出，文本情感分类的方法，不关心词语一级和句子一级的情感倾向，而是从整体上来分析整篇文档的情感倾向。

文本情感倾向计算是基于前面提到的词语级和句子级情感计算的基础上实现的，通常使用两种方法，一种是基于关键词统计的文本情感倾向计算方法，另一种是基于极性累加的文本情感倾向计算方法。

基于极性累加的文本情感倾向计算方法的主要思想是：利用前面介绍的某种句子级情感倾向性分析法，分析文本中出现的每一个句子，计算出句子倾向性及其强度，对于有倾向性的句子将强度累加，然后求平均值，将此作为文本的情感倾向信息；而对于没有极性的句子可将其强度记为 0。

基于关键词统计的文本情感倾向计算方法的主要思想是：首先给定情感词情感强度阈值 a、出现频度 b 和褒义词与贬义词的比值 e，然后计算从文档中抽取出的情感词的情感强度以及出现频度，如果它们的值分别大于给定的 a 和 b，则将这些词加入到关键词列表。最后对关键词列表进行统计，如果其中褒义词与贬义词相等，则文本是中性的，如果褒义词与贬义词的比值大于阈值 e，则文本的倾向被定义为褒义；否则文本的倾向定义为贬义。

3.1.5.3 色彩与情绪的关系

大自然的各种色彩使人产生各种感觉，并可陶冶人的情操。不同的颜色使人产生不同的情绪，从而引起人的心境发生变化。




心理学家对颜色与人的心理健康进行了研究。研究表明在一般情况下，红色表示快乐、热情，它使人情绪热烈、饱满，激发爱的情感。黄色表示快乐、明亮，使人兴高采烈，充满喜悦之情。绿色表示和平，使人的心里有安定、恬静、温和之感。蓝色给人以安静、凉爽、舒适之感，使人心胸开朗。灰色使人感到郁闷、空虚。黑色使人感到庄严、沮丧和悲哀。白色使人有素雅、纯洁、轻快之感。总之各种颜色都会给人的情绪带来一定的影响，使人的心理活动发生变化。

在本系统中，按照心理学中色彩和情绪的对应关系来展示微博中不同的心情。

下表列出部分本系统中情绪分类和对应颜色，完整见附录 A。

表 3-4 色彩-情绪对应表

分类	颜色	16 进制值	RGB 值
喜爱		#F76D02	247, 109, 2
愤怒		#000082	0, 0, 130
烦闷		#FF0000	255, 0, 0
悲伤		#3C8259	60, 130, 89

思		#8FBFE7	143, 191, 231
恐惧		#5F2688	95, 38, 136
惊奇		#FFFF00	255, 255, 0

### 3.1.6主流程

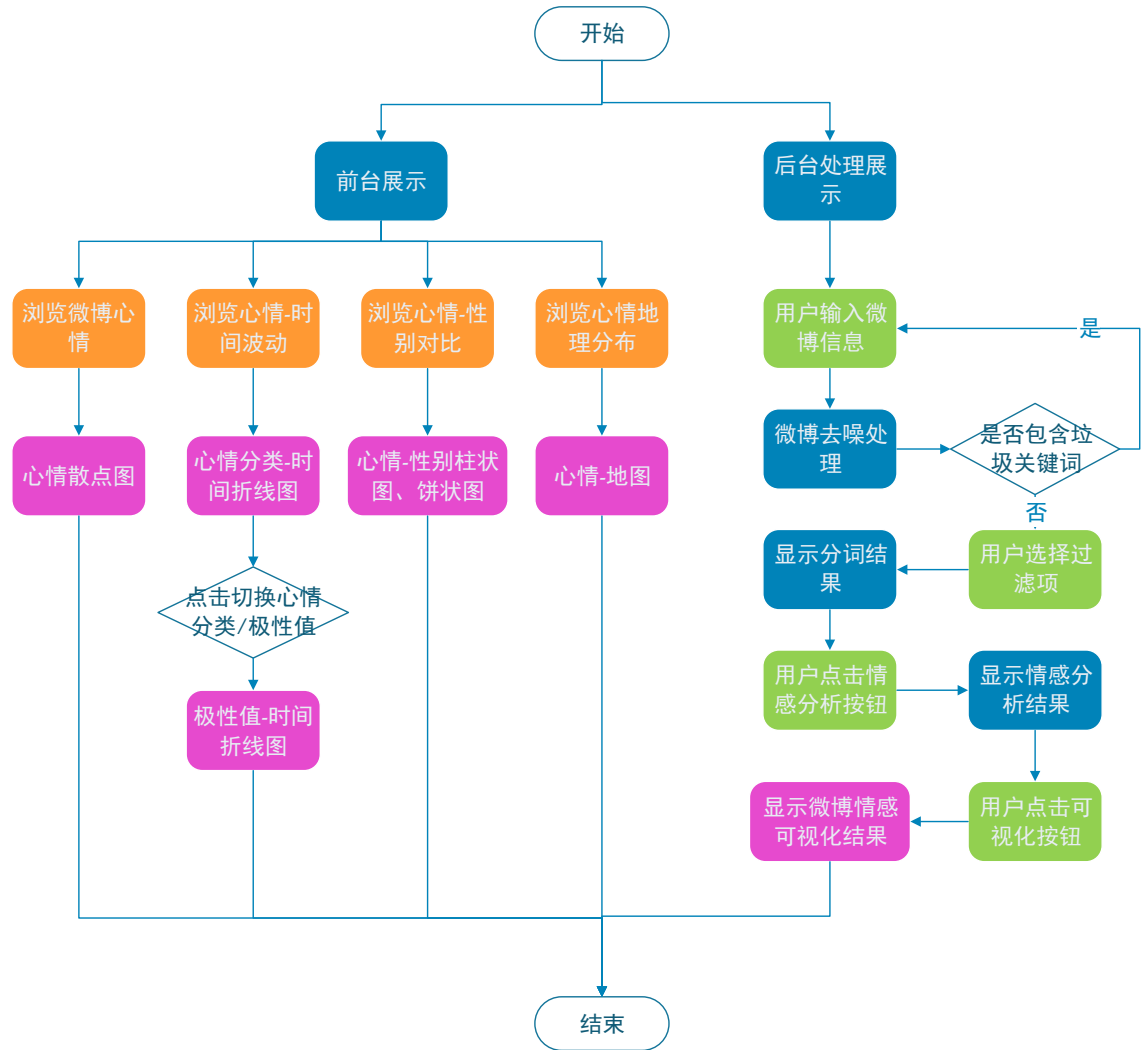


图 3-7 系统主流程图

从用户完整的使用过程来看，本系统的主要流程是：首先用户根据网址登录本系统的主界面。用户通过主界面的导航栏可以进入“浏览微博心情”，“浏览心情时间波动”，“浏览心情性别对比”“浏览心情地理分布”和“浏览热门话题”。用户可以通过点击详情，浏览具体的微博内容。可以点击情绪关键词，查看相关微博；亦可选择情感类别，浏览该种情感的微博。用户可以设置选定具体的心情分类，浏览不同心情所占比例的对比。用户可以通过点击选择是否隐藏或者显示特定的心情，展示心情随着时间的波动。用户可以点击查看热门话题。

### 3.1.7 系统实现方案

本系统分为“微博心情”展示模块和“微博热门话题”展示模块。图 2-5 体现了整个系统组件之间的通信过程和实现原理。

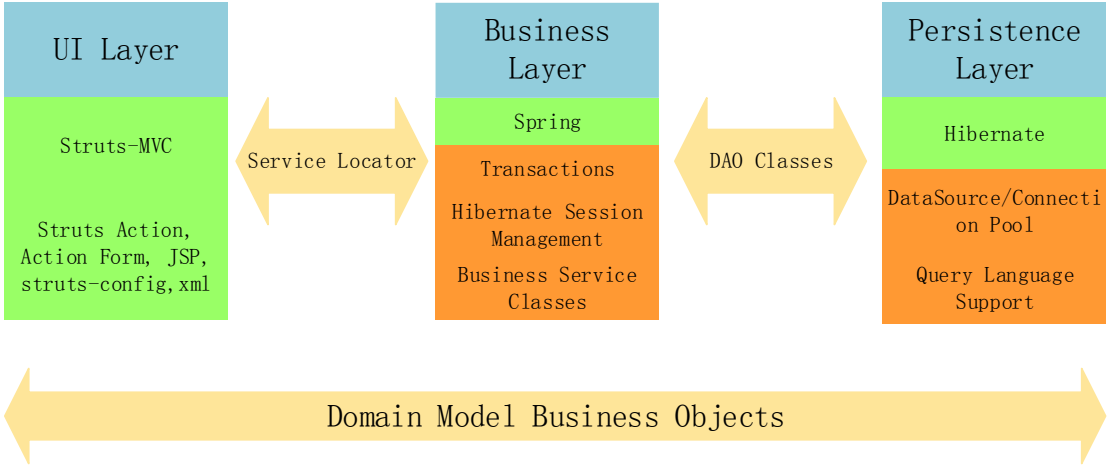


图 3-8 系统实现方案图

为了保证系统具有极佳的稳定性和良好的扩展性，并满足应用系统的基本要求，本项目选择了 J2EE 开发平台，为系统构建的整体技术架构。

下表具体描述了各个层级的具体细化描述：

表 3-5 系统层级划分图

层级	细化	描述
表现层		主要是 JSP 和 HTML 页面，用于接收用户的请求，以及返回操作数据，是应用程序访问的入口。
业务逻辑层		主要是对数据层进行操作，对数据逻辑层进行处理，如果数据层是积木，那么逻辑层就是堆积木的搭建
	Actions&Structs 2	MVC 架构的控制层。控制业务逻辑层与表现层的交互。
	Service 层	业务逻辑的实现。
数据访问层		主要是对原始数据的操作层，具体为业务逻辑层或表现层提供数据服务。
	DAO 层	负责数据访问对象与持久化对象的交互。
	持久化对象 Model	通过实体-关系映射 OR-Mapping 由数据库表得到对象
	数据库服务	对数据库进行操作

3.1.8功能需求与程序的关系

	心情散点展示	时间趋势分布	地理分布	性别比例分布	关键字展示	热门话题展示	关键字提取	情感分析	数据获取
心情散点展示	√						√	√	√
时间趋势分布		√					√	√	√
地理分布			√				√	√	√
性别比例分布				√			√	√	√
关键字展示					√		√		√
热门话题展示						√	√	√	√
关键字提取							√	√	√
情感分析								√	√
数据获取									√

3.1.9人工处理过程

3.1.9.1 建立停用词词典

在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉某些字或词，这些字或词即被称为 Stop Words（停用词）。

本小组以哈工大停用词表、四川大学机器智能实验室停用词库、百度停用词表为基础，进行删改后得到的一个较完整的停用词汇表，包含停用词 2231 条。

该停用词表主要用在处理自然语言文本之前过滤掉某些字或词，目的是缩小搜索范围以提高搜索性能。

表 3-6 停用词词典举例

词语	词语	词语	词语	词语
----	----	----	----	----



一些	不	与其说	为此	looking
一何	不尽	且不说	乃至	meanwhile
一转眼	不尽然	且说	二来	merely
上	不若	为止	云尔	might

### 3.1.9.2 建立程度词、否定词及关联词词表

程度词和否定词可统称为情感词的修饰词。对于情感词而言，否定词置反了其情感极性，而程度词（如非常、很、一般等）则加强或减弱了其情感极性。因而在进行情感分类时，必须考虑这两类词的影响。一般而言，在转折句中，往往转折词之后的语句才真正表达了作者的真实意图，故而在进行文本处理时，只需对转折词之后的语句进行分析。

程度词表以 Hownet 总结的中文程度级别词语为基础建立。在 HowNet 中，程度词共有 219 个，根据其修饰程度的强弱可分为六级：

- 一、“极其”级别。修饰程度最强。如“倍加”、“极度”等。
- 二、“很”级别。修饰程度很强。如“很”、“特别”等。
- 三、“较”级别。修饰程度较强。如“较为”、“愈来愈”等。
- 四、“稍”级别。修饰程度稍强。如“略微”、“一点儿”等。
- 五、“欠”级别。修饰程度欠强。如“相对”、“不怎么”等。
- 六、“超”级别。修饰程度过分强。如“过分”、“过于”等。

而否定词表与转折词表则为本小组自行收集所得。词表示例如表 3-6 所示。

表 3-7 词表示例

类别	程度	示例	数量
程度词	极其	百分之百、倍加备至、不得了不堪、不可开交……	69
	很	不少、颇为、甚、实在、太、太甚……	42
	较	大不了、多、更、更加、还、还要、较……	37
	稍	略微、略为、蛮、稍、稍稍、稍微……	29
	欠	半点、不大、不丁点儿、不甚、不怎么……	12
	超	过度、过分、过火、过劲、过了头……	30
否定词		不，弗，勿，毋，未，非，微，无……	12
转折词		但，但是，可是，只是，然而……	9
假设词		如果，假设，假若，假如，如若，倘若，倘使……	18

### 3.1.9.3 部分数据源的获取

- 垃圾微博的获取采用人工收集法，本项目中一共收集了 **805** 条垃圾微博，经过分词、停用词过滤后得到了 **62** 个噪声词。
- 垃圾微博举例：
  - ✧ RC **套装**来袭。两片 RC 面膜+一瓶 RC 精华素，RC 精华素超补水，能修复过敏，减淡痘印淡化斑点！
  - ✧ 转**代理** 老公对一切黏糊糊的东西都是超级抗拒的[大哭]**面膜**，唇膏更是拒之千里[难过]我做面膜时他都离的远远的就怕粘到他身上了[撇嘴]没想到这次居然愿意做 KT **面膜**[惊讶][惊讶][惊讶]好惊讶！实属难得，做出来的效果也很惊讶[撇嘴][撇嘴][惊讶][惊讶]好[吐]。
  - ✧ **拔草**啦~~~超爱的眼霜啊!!! 不似大牌般死贵，但是性价比超高~~纯中药粉，完全达到口服的标准。经常熬夜熬出的黑眼圈，用了这个眼霜，竟然一周就消下去了~~还有针对眼袋、眼纹等问题的眼部产品！经常坐在电脑前，或者夜猫族滴 MM，一定要尝试下！分享>> [神马]<http://t.cn/RPUDWN4>
  - ✧ **福利** NO.18[禮物]转发加关注@黄秋媚 yo 8 月 23 日抽出 2 名幸运儿**包邮**送以下美衣一件。
- 噪声词举例：套装、淘宝、代理、面膜、福利、拔草、包邮

## 3.2 数据库设计

### 3.2.1 概述

本系统采用 MySQL5.6 数据库，轻量、快速，且功能强大，移植性较好。

系统数据库设计遵循如下原则：

#### 1) 命名规范

命名项	命名规范	备注
表名	数据表名称必须以有特征含义的单词或缩写组成，中间以用大写分割词语。表名称不能用双引号包含	关系表的表名为“表名 1 + to + 表名 2”的形式，或表名 1/2 采用缩写的形式
字段名	主键都设为 ID，其他字段名称必须用字母开头，采用有特征含义的单词或缩写，不能用双引号包含	外键名为被引用的键所在表名/缩写+Id

#### 2) 数据类型

数据类型	选择准则	备注
------	------	----

字符型	固定长度的字符串类型采用 CHAR，长度不固定的字符串类型采用 VARCHAR。避免在长度不固定的情况下采用 CHAR 类型	如果在数据迁移时出现以上情况，则必须使用 trim() 函数截去字符串后的空格
数字型	数字型字段尽量采用 INT 类型	默认长度为 11
日期和时间	由数据导入或外部应用程序产生的日期时间类型采用 DATETIME 类型	无

### 3.2.2 逻辑结构设计

用户在使用本系统进行相应的功能操作时，主要是通过前台 Web 端或者移动设备端进行相应功能的操作，例如查看微博心情分布，查看热门话题等，前端将相关信息提取出来后，发送到后台服务器中，而后台服务器则对用户的相关信息进行存储或处理。

下图是本系统的逻辑结构设计图：

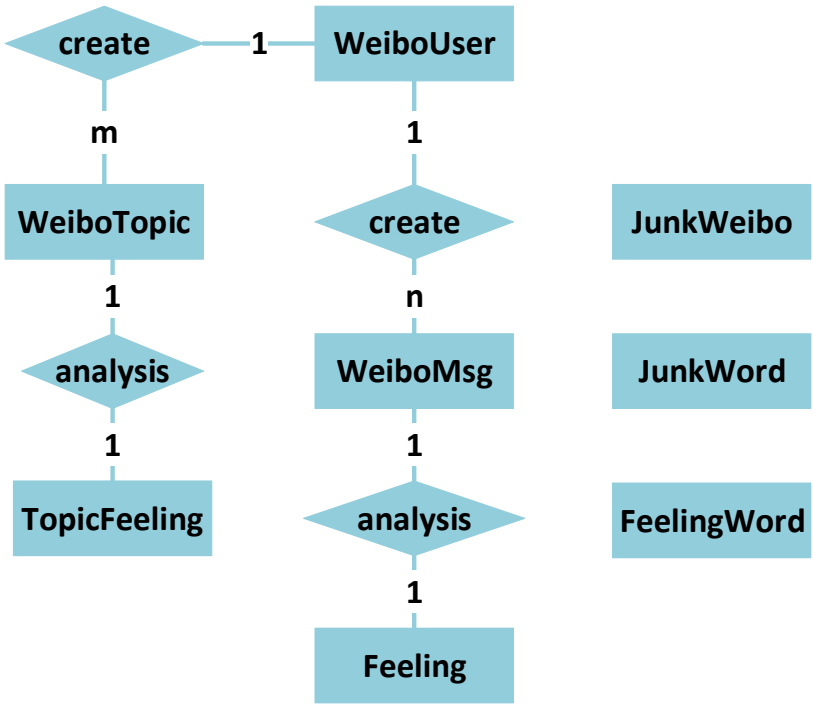


图 3-9 系统逻辑结构设计图

数据实体说明如表 3-8 所示。

数据实体	说明
weibouser	微博用户的个人信息，包含用户 id，用户名，用户性别等。
hotuser	热门微博用户的个人信息，包含用户 id，用户名，用户性别等。
weibomsg	使用微博 API 获取的微博消息，包含消息 id，消息内容，消息发布者 id 等。

hotusermsg	热门微博用户的微博消息。
weibotopic	微博话题对应的评论信息,包含消息 id, 消息内容, 消息发布者 id, 话题关键字等。
feeling	微博的情感分析信息, 包含情感分类, 情感描述, 情感极性值等
hotuserfeeling	热门微博用户的情感分析信息。
topicfeeling	微博话题的情感分析信息, 包含情感分类话题关键字, 微博影响力等。
feelingword	情感词记录, 包含情感词, 情感词强度, 情感词极性 etc。
junkweibo	提取出的垃圾微博, 包含微博内容, 微博垃圾关键词等。
junkword	筛选出的微博垃圾词。

表 3-8 数据实体说明

3.2.3物理结构设计

给出本系统内所使用的每个数据结构中的每个数据项的存储要求，访问方法、存取单位、存取的物理关系（索引、设备、存储区域）、设计考虑和保密条件。

以下是本系统的物理结构设计：

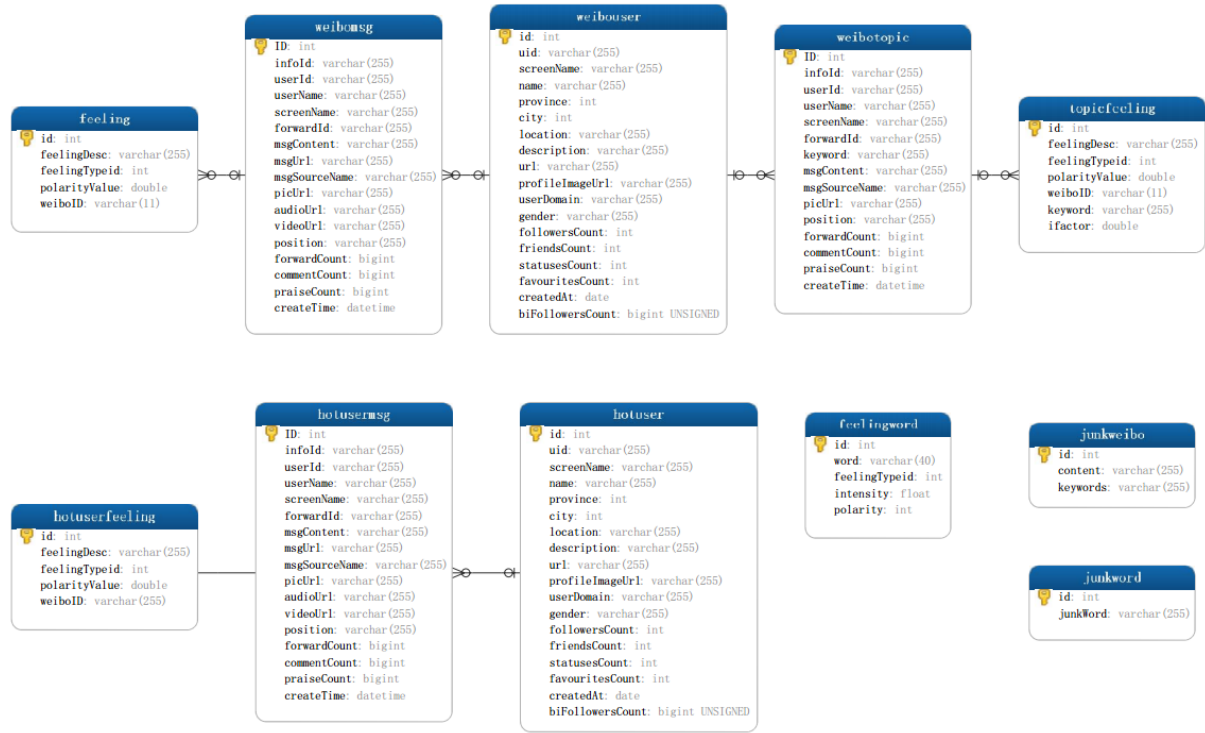


图 3-10 系统物理结构设计图

各数据表具体定义如下：

表 3-9 weibomsg 数据表

weibomsg				
字段	类型	空	默认	注释
<u>ID</u>	int(11)	否	1	主键
infoId	varchar(255)	是		微博 ID
userId	varchar(255)	否		用户 ID
userName	varchar(255)	是		用户名
screenName	varchar(255)	是		屏幕名
forwardId	varchar(255)	是		转发微博 ID
msgContent	varchar(255)	是		微博内容
msgUrl	varchar(255)	是		微博 URL
msgSourceName	varchar(255)	是		微博来源
picUrl	varchar(255)	是		图片 URL
audioUrl	varchar(255)	是		音频 URL
videoUrl	varchar(255)	是		视频 URL
position	varchar(255)	是		地理位置
forwardCount	bigint(16)	是	0	转发数
commentCount	bigint(16)	是	0	评论数
praiseCount	bigint(16)	是	0	赞数
createTime	datetime	是	NULL	发布时间

表 3-10 weibuser 数据表

weibouser				
字段	类型	空	默认	注释
<u>id</u>	int(11)	否		主键
uid	varchar(255)	否		用户 UID
screenName	varchar(255)	是		用户昵称
name	varchar(255)	是		友好显示名称
province	int(11)	是	0	省份编码(参考省份编码表)
city	int(11)	是	0	城市编码(参考城市编码表)
location	varchar(255)	是		地址
description	varchar(255)	是		个人描述

url	varchar(255)	是		用户博客地址
profileImageUrl	varchar(255)	是		自定义图像
userDomain	varchar(255)	是		用户个性化 URL
gender	varchar(255)	是		性别, m—男, f—女, n—未知
followersCount	int(11)	是	0	粉丝数
friendsCount	int(11)	是	0	关注数
statusesCount	int(11)	是	0	微博数
favouritesCount	int(11)	是	0	收藏数
createdAt	date	是	NULL	创建时间
biFollowersCount	bigint(20)	是	0	互粉数

表 3-11 weibotopic 数据表

weibotopic				
字段	类型	空	默认	注释
<u>ID</u>	int(11)	否		主键
infoId	varchar(255)	是		微博 ID
userId	varchar(255)	是		用户 ID
userName	varchar(255)	是		用户名
screenName	varchar(255)	是		屏幕名
forwardId	varchar(255)	是		转发微博 ID
keyword	varchar(255)	是		微博属于的话题关键字
msgContent	varchar(255)	是		微博内容
msgSourceName	varchar(255)	是		微博来源
picUrl	varchar(255)	是		图片 URL
position	varchar(255)	是		地理位置
forwardCount	bigint(16)	是	0	转发数
commentCount	bigint(16)	是	0	评论数
praiseCount	bigint(16)	是	0	赞数
createTime	datetime	是	NULL	发布时间

表 3-12 feeling 数据表

feeling				
字段	类型	空	默认	注释
<u>ID</u>	int (11)	否	1	主键
feelingDesc	varchar (255)	是	NULL	心情匹配关键词
feelingTypeid	int (11)	是	NULL	心情分类 1-23
polarityValue	double (11, 4)	是	NULL	极性值
weiboID	int (11)	是	NULL	外键

表 3-13 topicfeeling 数据表

topicfeeling				
字段	类型	空	默认	注释
<u>ID</u>	int (11)	否		主键
feelingDesc	varchar (255)	是		心情匹配关键词
feelingTypeid	int (11)	是	NULL	心情分类 1-23
polarityValue	double (11, 4)	是	NULL	极性值
weiboID	int (11)	是	NULL	外键
keyword	varchar (255)	是		所属的话题名称
ifactor	double (16, 4)	是	NULL	微博影响力因数

表 3-14 feelingword 数据表

feelingword				
字段	类型	空	默认	注释
<u>ID</u>	int (11)	否		主键
word	varchar (40)	否		情感词
feelingTypeid	int (2)	否		情感分类序号
intensity	float (4, 1)	否		情感强度 , 1.0-3.0
polarity	int (4)	否		情感极性, 1, 0, -1

表 3-15 junkweibo 数据表

junkweibo				
-----------	--	--	--	--

字段	类型	空	默认	注释
<u>ID</u>	int(11)	否		主键
content	varchar(255)	否		微博内容
keywords	varchar(255)	是	NULL	微博关键字

表 3-16 junkword 数据表

junkword				
字段	类型	空	默认	注释
ID	int(11)	否		主键
junkWord	varchar(255)	是	NULL	垃圾词

表 3-17 hotuser 数据表

weibouser				
字段	类型	空	默认	注释
<u>id</u>	int(11)	否		主键
uid	varchar(255)	否		用户 UID
screenName	varchar(255)	是		用户昵称
name	varchar(255)	是		友好显示名称
province	int(11)	是	0	省份编码(参考省份编码表)
city	int(11)	是	0	城市编码(参考城市编码表)
location	varchar(255)	是		地址
description	varchar(255)	是		个人描述
url	varchar(255)	是		用户博客地址
profileImageUrl	varchar(255)	是		自定义图像
userDomain	varchar(255)	是		用户个性化 URL
gender	varchar(255)	是		性别, m--男, f--女, n--未知
followersCount	int(11)	是	0	粉丝数
friendsCount	int(11)	是	0	关注数
statusesCount	int(11)	是	0	微博数
favouritesCount	int(11)	是	0	收藏数



createdAt	date	是	NULL	创建时间
biFollowersCount	bigint(20)	是	0	互粉数

表 3-18 hotusermsg 数据表

weibotopic				
字段	类型	空	默认	注释
<u>ID</u>	int(11)	否		主键
infoId	varchar(255)	是		微博 ID
userId	varchar(255)	是		用户 ID
userName	varchar(255)	是		用户名
screenName	varchar(255)	是		屏幕名
forwardId	varchar(255)	是		转发微博 ID
keyword	varchar(255)	是		微博属于的话题关键字
msgContent	varchar(255)	是		微博内容
msgSourceName	varchar(255)	是		微博来源
picUrl	varchar(255)	是		图片 URL
position	varchar(255)	是		地理位置
forwardCount	bigint(16)	是	0	转发数
commentCount	bigint(16)	是	0	评论数
praiseCount	bigint(16)	是	0	赞数
createTime	datetime	是	NULL	发布时间

表 3-19 hotuserfeeling 数据表

feeling				
字段	类型	空	默认	注释
<u>ID</u>	int(11)	否	1	主键
feelingDesc	varchar(255)	是	NULL	心情匹配关键词
feelingTypeid	int(11)	是	NULL	心情分类 1-23
polarityValue	double(11,4)	是	NULL	极性值
weiboID	int(11)	是	NULL	外键

### 3.2.4 数据结构与程序的关系

服务器程序在对数据进行展示的时候，需对数据库数据结构，也就是数据表进行查询：在数据展示的过程中都需要对数据库中的所有表，进行联合查询。

物理数据结构主要用于各模块之间函数的信息传递。接口传递的信息将是以数据结构封装了的数据，以参数传递或返回值的形式在各模块间传输。出错信息将送入显示模块中。从微博提供的接口获取到的数据，在经过处理分析后封装入特定的数据结构内，持久化入数据库。

关系表示如下图：

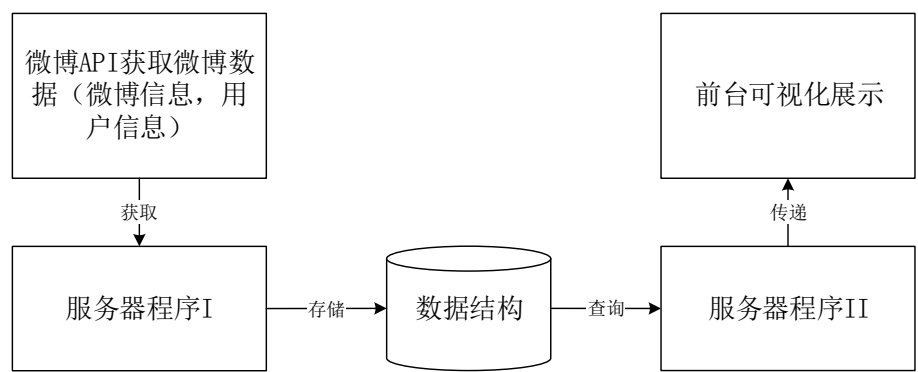


图 3-11 数据结构与程序关系图

### 3.3 关键算法设计

#### 3.3.1 垃圾微博过滤算法

微博平台上除了普通用户自己分享的观点看法，还存在许多商业营销和广告信息，此类信息在本系统中被划为垃圾微博，不在有效信息范围之内，需要事先进行过滤。其过滤算法主要分为两个步骤：建立噪声词库、垃圾微博的过滤，具体流程如图 3-12 所示：

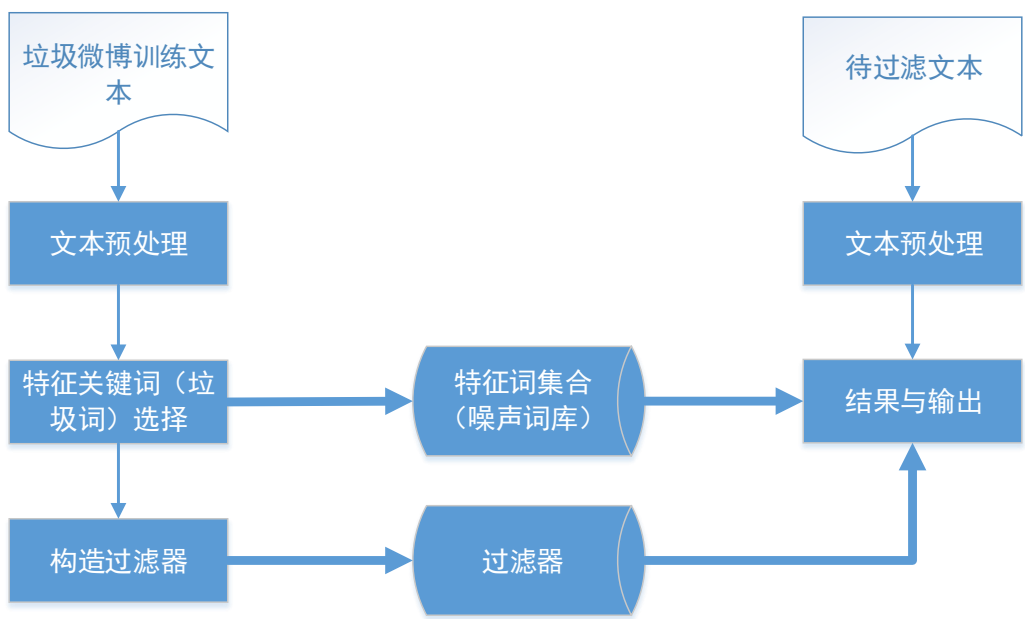


图 3-12 垃圾微博过滤流程图

### 3.3.2情感分类算法

微博作为一种细粒度的文本，具有其自身特点，文本短小，表达方式活泼，传统的文本情感倾向性研究方法大多已不适用，因此在本系统中，针对微博的特点，主要采用基于情感词典和语义规则的方法进行微博的情感倾向性分析及情感分类。微博文本情感的主要识别过程共分两个步骤：

1) 建立情感模型，其主要流程如下：

第一步：将微博文本  $T$  进行预处理，包括话题、用户名、链接、特殊符号等无效信息的过滤，并对微博文本  $T$  进行分类，将其拆分为原创文本  $T_o$  及转发文本  $T_f$  两部分。

第二步：对文本进行分词，根据词性过滤等提取出候选的情感词集合。

第三步：将候选词与情感词典匹配，得到文本中包含的情感词集合  $W = \{w_1, w_2, \dots, w_n\}$ 。

第四步：根据语义规则，遍历情感词集合，提取出该集合中情感词的修饰词（程度副词集合  $E$  及否定词集合  $D$ ），构建相应的情感模型  $M_i = \{w_i, D, E\}$ 。情感模型构建流程如图 3-13 所示。

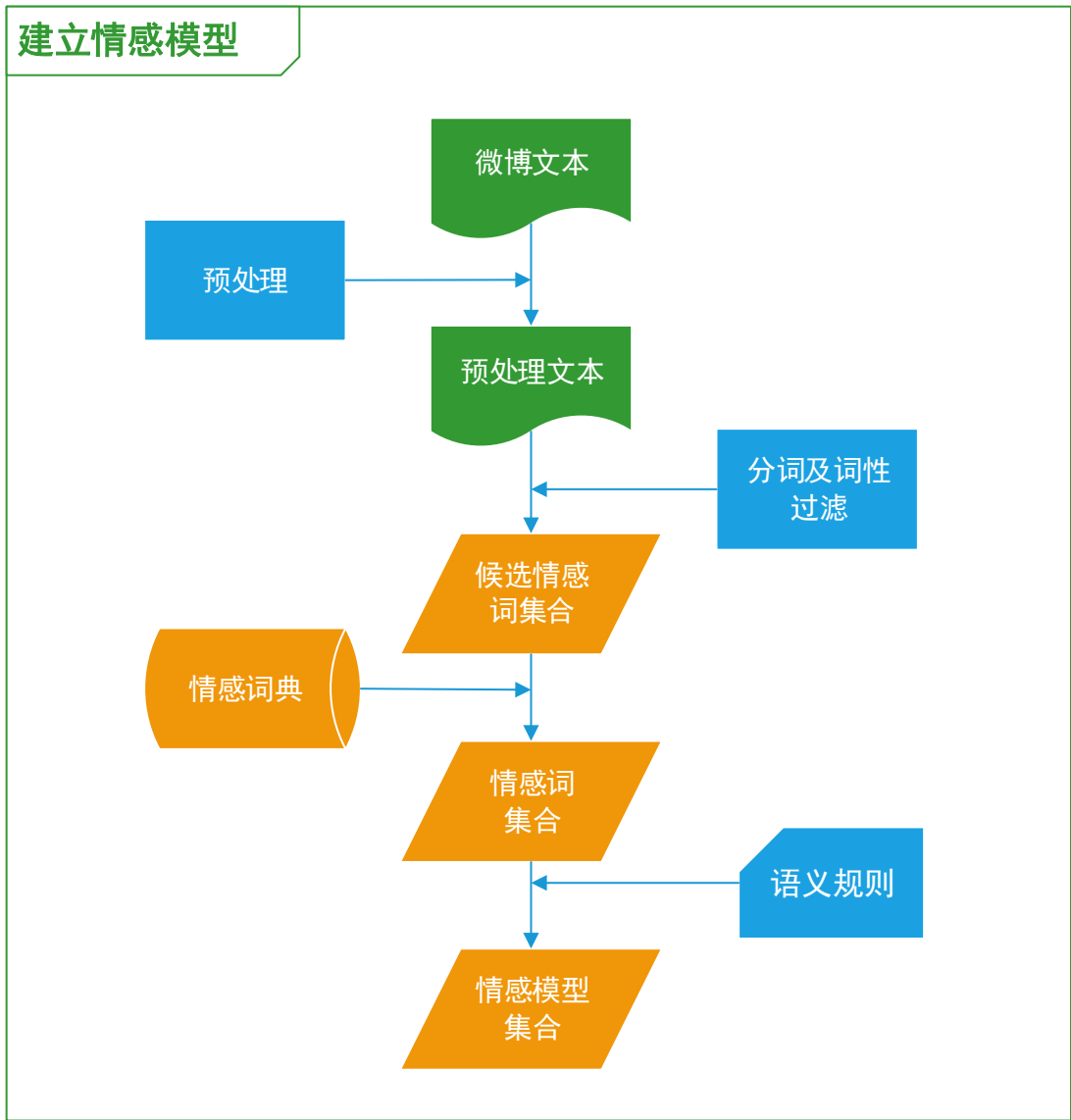


图 3-13 情感模型建立流程图

2) 根据情感模型进行文本情感分类及情感倾向性判断，其主要流程如下：

第一步：遍历情感模型集合  $M$ ，根据情感模型  $M_i = \{w_i, D, E\}$ ，计算该模型的情感极性值

$Polarity(M_i)$ ，其计算公式如公式 3-1 所示。各参数定义如表 3-16 所示。

$$Polarity(M_i) = \delta(w_i) * \gamma(w_i) * \eta(w_i) \quad (\text{公式 3-1})$$

表 3-20 情感极性值计算公式参数定义

参数	计算公式	说明
情感词极性值 $\delta(w_i)$	$\delta(w_i) = R(w_i) * I(w_i)$	$R(w_i)$ 为该词极性， $I(w_i)$ 为该词强度
程度副词影响因数 $\gamma(w_i)$	$\gamma(w_i) = \prod_{k=1}^m Deg(d_k)$	$Deg(d_k)$ 为该程度词情感扩大倍数
否定词影响因数 $\eta(w_i)$	$\eta(w_i) = (-1)^n$	n 为否定词频数

第二步：计算情感模型  $M_i$  中情感词  $w_i$  的权重  $Weight(w_i)$ ，其计算方法如公式 3-2 所示。

$$Weight(w_i) = |Polarity(M_i)| \quad (\text{公式 3-2})$$

第三步：根据情感词集合  $W$  中各个情感词的权重，遍历累加出各情感分类的权值  $Weight(ft_i)$ ，排序选出权值最高的情感分类，该情感分类即为微博文本的情感分类  $ft(T)$ 。

$$Weight(ft_i) = \sum_{j=1}^n Weight(w_j), \quad ft(w_j) \in ft_i \quad (\text{公式 3-3})$$

$$ft(T) = \max\{Weight(ft_i)\} \quad (\text{公式 3-4})$$

第四步：将所有情感模型的极性值累加，得出微博文本的情感极性值。其计算方法如公式 3-3 所示。

$$Polarity(T) = \alpha * \sum_{i=1}^k Polarity(M_i) + (1 - \alpha) * \sum_{i=k+1}^n Polarity(M_j) \quad (\text{公式 3-3})$$

其中， $T = \{T_o, T_F\}$ ， $M_i \in T_o$ ， $M_j \in T_F$ ， $\alpha$  为原创微博影响权数。

情感倾向判断及分类流程图如图 3-14 所示。

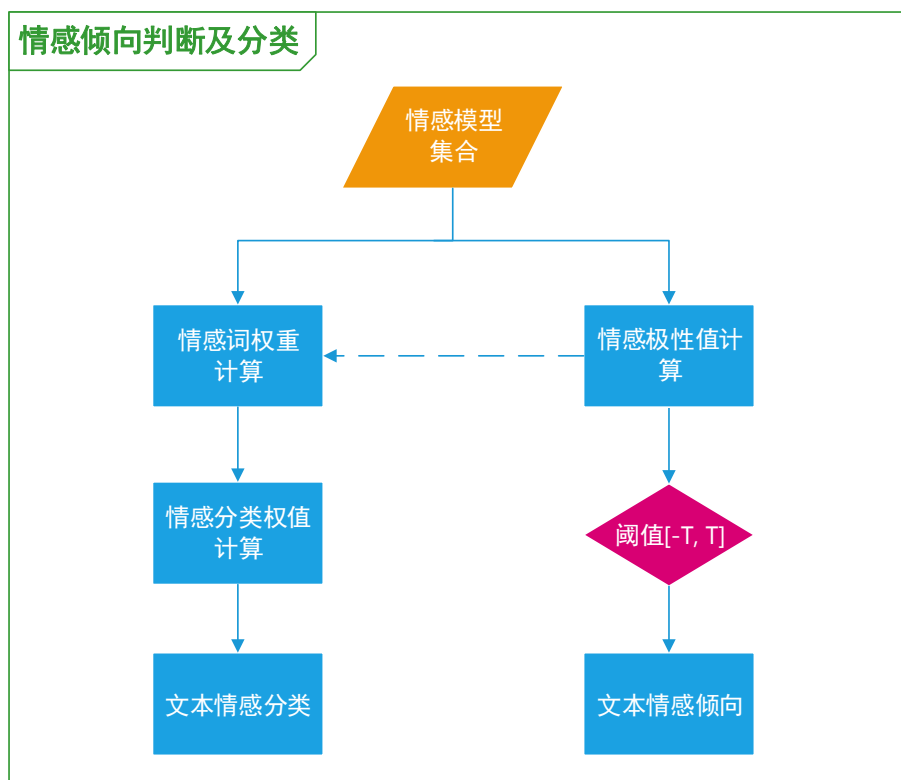


图 3-14 情感分类处理流程图

### 3.3.3 微博影响力评估算法

微博现在已经成为热点事件产生和谈论的重要场所，在进行话题情感分析的时候，除了考虑各个参与者本身的情感表达，信息本身的吸引力在网络中的传播同样重要。高影响力微博的存在和转发是引起信息持续传播和形成更大传播规模的关键因素因而，在微博话题分析中，本系统加入了微博影响力作为一个重要因数，作为整个话题情感分类的判断信息。

微博影响力主要通过微博的转发数、评论数和赞数体现。微博的转发数、评论数和赞数越多，说明微博内容越受关注，同时影响力也越大，根据微博、评论数和赞数对微博影响力贡献大小的不同，得出微博内容影响力计算方法，如公式 2-4 所示。

$$MI = \sqrt[3]{MF} + \sqrt{MC} + \sqrt{MP} \quad (\text{公式 3-4})$$

其中，MI 为该条微博内容影响力因数，MF 为该条微博转发数，MC 为该条微博评论数，MP 为该条微博赞数。

### 3.3.4 关键字提取算法

为了进一步挖掘微博热门话题的潜在信息，提取其评论关键字是最为直接也十分必要的工作，因此我们设计了一套关键字的提取算法，其主要处理过程如图 3-15 所示：

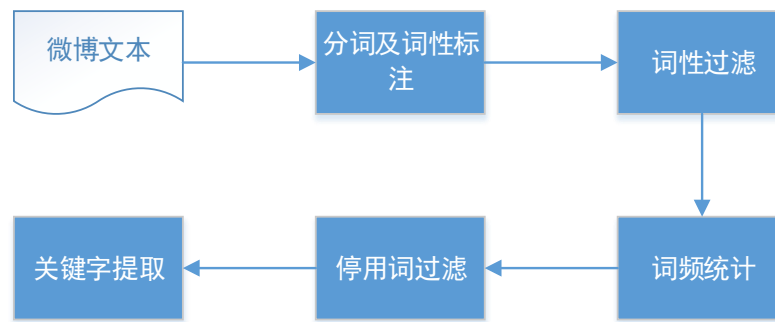


图 3-15 关键字提取流程图

# 4. 详细设计

该部分是基于概要设计编写的详细设计说明书。详细描述了本项目的框架、模块划分、模块结构和模块设计。项目的开发基于此说明书进行。

## 4.1 包设计

本系统中包含了大量的类，因此也会存在大量的事物，为了能更加有效地进行整合、生产宏观模型，就需要对其分组，就是本系统中的包设计

下图是本系统中的包设计、包间的依赖关系、包中类设计：

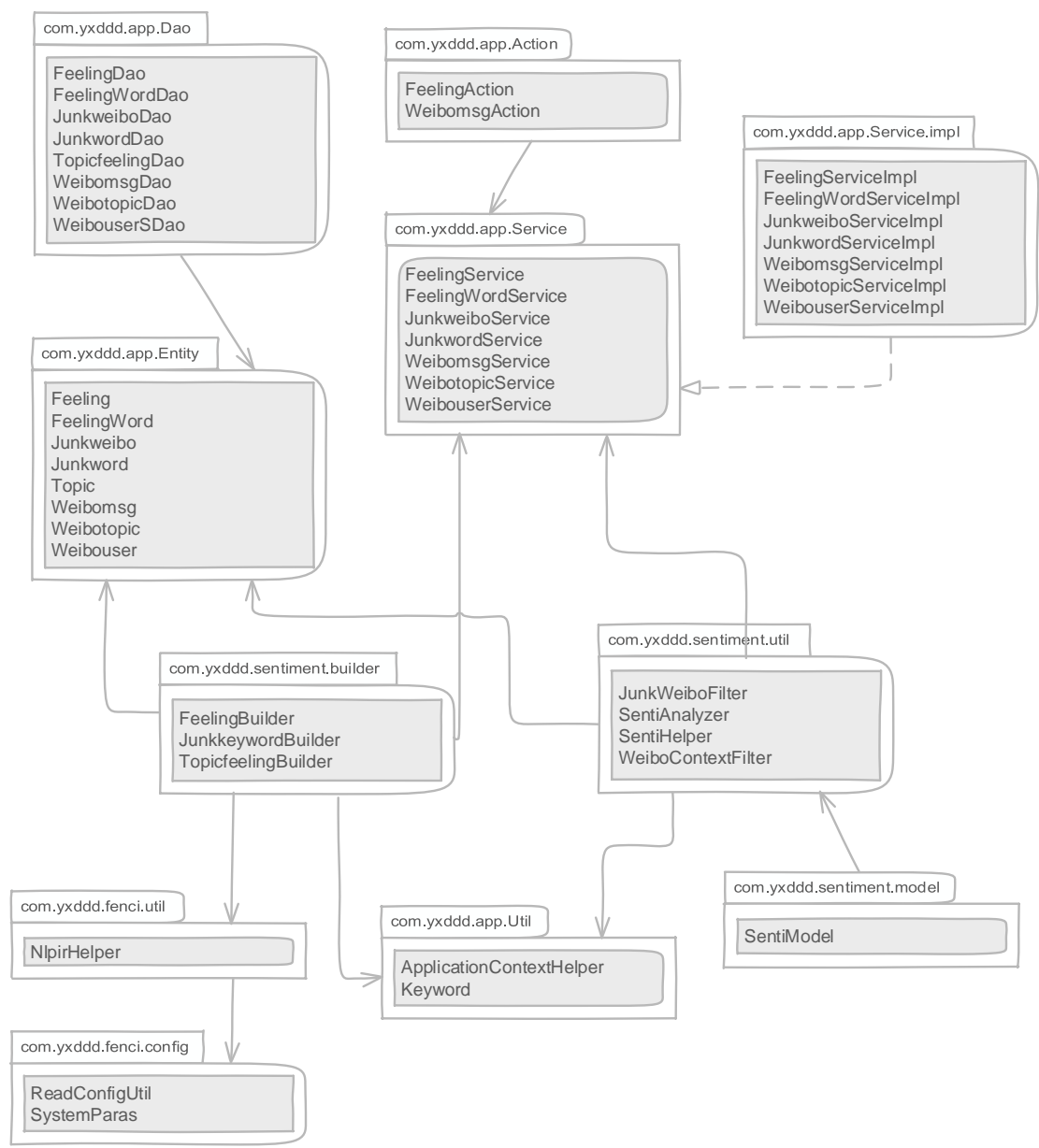


图 4-1 项目包图

下表是本系统中的各个包的设计说明：

表 4-1 包设计说明表

包名	设计说明
<code>com.yxddd.app.Entity</code>	该包主要存放数据库表中所对应的实体类。
<code>com.yxddd.app.Dao</code>	主要负责数据访问, 该包中的类主要封装了对数据库的访问 (只包含最原子的数据操作), 供高层调用。 依赖于 <code>com.yxddd.app.Entity</code> 。
<code>com.yxddd.app.Service</code>	该包中主要存放了高层调用 Dao 的接口定义。
<code>com.yxddd.app.Service.impl</code>	实现了 <code>com.yxddd.app.Service</code> 中所定义的所有接口。 调用 <code>com.yxddd.app.Dao</code> 实现数据的访问。 包含一部分的业务逻辑。 依赖于 <code>com.yxddd.app.Dao</code> 和 <code>com.yxddd.app.Entity</code> 。
<code>com.yxddd.app.Action</code>	调用 Service 以完成业务逻辑。 依赖于 <code>com.yxddd.app.Service</code> , <code>com.yxddd.app.Entity</code> 。
<code>com.yxddd.fenci.config</code>	存放了中文分词工具的配置类。
<code>com.yxddd.fenci.util</code>	存放了中文分词工具所提供的一些公共类。 例如加载配置文件、定义中文分词所需接口、分词、获取词性标注等实现类。 依赖 <code>com.yxddd.fenci.config</code> 。
<code>com.yxddd.sentiment.model</code>	存放情感模型所对应的实体类。
<code>com.yxddd.app.util</code>	存放 Java Application 工具类。 负责在加载 Java Application 程序中读取 Spring 的配置文件, 使用 Spring 注入对象。
<code>com.yxddd.sentiment.builder</code>	存放各种词典、词库的生成器。 依赖 <code>com.yxddd.app.util</code> <code>com.yxddd.app.Entity</code> 、 <code>com.yxddd.app.Service</code> 。
<code>com.yxddd.sentiment.util</code>	存放所有处理情感分析的类。 依赖 <code>com.yxddd.app.Entity</code> 、 <code>com.yxddd.app.Service</code> 。

## 4.2 类设计

本系统共有 11 个包, 46 个类。本小节以包为单位, 对每个包所含有的类及类与类间的调用关系进行说明。

表 4-2 app.entity 相关类设计

com.yxddd.app.entity 相关类	
类名	设计说明



Weibouser	微博用户的个人信息实体类，对应于数据库中的 weibouser 表。
Weibomsg	微博消息实体类，对应于数据库中的 weibomsg 表。
Weibotopic	微博话题对应的评论信息实体类，对应于数据库中的 weibotopic 表。
Feeling	微博的情感分析信息实体类，对应于数据库中的 feeling 表。
Topicfeeling	微博话题的情感分析信息实体类，对应于数据库中的 topicfeeling 表。
Feelingword	情感词实体类，对应于数据库中的 feelingword 表。
Junkweibo	垃圾微博实体类，对应于数据库中的 junkweibo 表。
Junkword	微博垃圾词实体类，对应于数据库中的 junkword 表。

表 4-3 app.dao 相关类设计

com.yxddd.app.dao 相关类	
类名	设计说明
WeibouserDao	封装与 Weibouser 实体类相关的数据访问方法。
WeibomsgDao	封装与 Weibomsg 实体类相关的数据访问方法。
WeibotopicDao	封装与 Weibotopic 实体类相关的数据访问方法。
FeelingDao	封装与 Feeling 实体类相关的数据访问方法。
TopicfeelingDao	封装与 Topicfeeling 实体类相关的数据访问方法。
FeelingwordDao	封装与 Feelingword 实体类相关的数据访问方法。
JunkweiboDao	封装与 Junkweibo 实体类相关的数据访问方法。
JunkwordDao	封装与 Junkword 实体类相关的数据访问方法。

表 4-4 app.service 相关类设计

com.yxddd.app.service 相关类	
类名	设计说明
WeibouserService	定义了与 Weibouser 实体类相关的业务逻辑接口
WeibomsgService	定义了与 Weibomsg 实体类相关的业务逻辑接口
WeibotopicService	定义了与 Weibotopic 实体类相关的业务逻辑接口
FeelingService	定义了与 Feeling 实体类相关的业务逻辑接口
TopicfeelingService	定义了与 Topicfeeling 实体类相关的业务逻辑接口
FeelingwordService	定义了与 Feelingword 实体类相关的业务逻辑接口
JunkweiboService	定义了与 Junkweibo 实体类相关的业务逻辑接口

JunkwordService	定义了与 Junkword 实体类相关的业务逻辑接口
-----------------	----------------------------

表 4-5 app.service.impl 相关类设计

com.yxddd.app.service.impl 相关类	
类名	设计说明
WeibouserServiceImpl	实现了 WeibouserService 接口，通过调用 WeibouserDao 实现与 Weibouser 相关的数据访问。
WeibomsgServiceImpl	实现了 WeibomsgService 接口，通过调用 WeibomsgDao 实现与 Weibomsg 相关的数据访问。
WeibotopicServiceImpl	实现了 WeibotopService 接口，通过调用 WeibotopDao 实现与 Weibotop 相关的数据访问。
FeelingServiceImpl	实现了 FeelingService 接口，通过调用 FeelingDao 实现与 Feeling 相关的数据访问。
TopicfeelingServiceImpl	实现了 TopicfeelingService 接口，通过调用 TopicfeelingDao 实现与 Topicfeeling 相关的数据访问。
FeelingwordServiceImpl	实现了 FeelingwordService 接口，通过调用 FeelingwordDao 实现与 Weibouser 相关的数据访问。
JunkweiboServiceeeImpl	实现了 JunkweiboService 接口，通过调用 JunkweiboDao 实现与 Junkweibo 相关的数据访问。
JunkwordServiceeeImpl	实现了 JunkwordService 接口，通过调用 JunkwordDao 实现与 Junkword 相关的数据访问。

表 4-6 app.action 相关类设计

com.yxddd.app.action 相关类	
类名	设计说明
WeibomsgAction	定义了响应情感统计界面不同点击事件的不同方法，根据事件的不同调用 WeiboService 的相应方法获取统计数据，以 json 格式传至前台。
FeelingAction	定义了响应情感分析处理界面不同点击及输入事件的方法，调用情感处理模块中的 JunkwordFilter、SentiAnalyzer 实现对用户输入文本的情感分析及可视化。

表 4-7 fenci.config 相关类设计

com.yxddd.fenci.config 相关类	
类名	设计说明
ReadConfigUtil	分词工具 Nlpir 类库配置文件 nlpir.properties 的加载类。

<b>SystemParams</b>	分词工具 Nlpir 类库配置参数加载类。
---------------------	-----------------------

表 4-8 fenci.util 相关类设计

com.yxddd.fenci.util 相关类	
类名	设计说明
<b>NlpirHelper</b>	分词辅助工具类。通过调用 ReadConfigUtil 读取分词工具 Nlpir 类库的配置信息以调用分词器。用于将传入的文本进行分词。

表 4-9 sentiment.model 相关类设计

com.yxddd.sentiment.model 相关类	
类名	设计说明
<b>SentiModel</b>	情感模型实体类。用于存放情感词词语、情感词的位置、情感词的修饰词及句式信息。

表 4-10 app.util 相关类设计

com.yxddd.app.util 相关类	
类名	设计说明
<b>ContextApplicationHelper</b>	Java Application 工具类。负责在加载 Java Application 程序中读取 Spring 的配置文件，使用 Spring 注入对象。

表 4-11 sentiment.builder 相关类设计

com.yxddd.sentiment.builder 相关类	
类名	设计说明
<b>FeelingBuilder</b>	微博情感分析信息实体（Feeling）的构造器。先调用 I/O 模块的 WeibomsgService 获取全部微博信息，通过 WeiboContentFilter 与 SeniAnalyzer 来进行微博文本分析以构造 Feeling 实体，并将之通过 I/O 模块的 FeelingService 存入数据库中。
<b>TopicfeelingBuilder</b>	微博话题情感分析信息实体（Topicfeeling）的构造器。通过调用 WeiboContentFilter 与 SeniAnalyzer 来进行微博文本分析以构造 Topicfeeling 实体，并根据微博影响力评估算法计算话题评论的影响力，将之通过 I/O 模块的 TopicfeelingService 存入数据库中。
<b>JunkKeywordBuilder</b>	微博垃圾词实体（Junkword）的构造器。通过调用 I/O 模块的 JunkweiboService 获取所有垃圾微博（Junkweibo），调用文本处理模块的 NlpirHelper 对微博分词，提取出频数达到一定阈值的垃圾关键字作为垃圾词，并通过 JunkwordService 存入数据库中。

表 4-12 sentiment.util 相关类设计

com.yxddd.sentiment.util 相关类	
类名	设计说明
SeniAnalyzer	微博情感分析器。通过调用文本处理模块的 NlpirHelper 获取文本分词结果，词性过滤后获取候选情感词，再通过调用 I/O 模块的 FeelingwordService 来进行匹配获取情感词，根据语义规则调用 SentiHelper 构建相应情感模型（SentiModel），根据情感分类算法获取微博的对应情感分类和极性值。
SentiHelper	微博情感分析辅助工具类。存有微博程度词、否定词、转折词及情感类别信息。可以其判断词语是否属于情感修饰词。
WeiboContentFilter	微博文本过滤器。用于对文本进行简繁转换及无效信息过滤。
JunkWeiboFilter	垃圾微博过滤器。通过调用 I/O 模块的 WeibomsgSevice 与 JunkwordSevice 获取所有微博信息与垃圾词信息，将含有垃圾词的信息标记为垃圾微博并删除。

### 4.3 模块设计

本系统共分为 5 个功能模块，其中，情感分析模块、噪声过滤模块与可视化模块属于高层模块，I/O 模块与文本处理模块属于底层模块。由于高层模块均需通过调用底层模块实现其功能，故而两个底层模块设计不再单独给出。本节将介绍系统的三大高层模块及每个高层模块模块所调用的部分底层模块信息。

#### 4.3.1 情感分析模块

情感分析模块用于微博文本的情感分类及倾向性判断。首先需调用 I/O 模块获取微博语料，通过文本处理模块对语料进行预处理后得到可供分析的有效文本，对有效文本进行情感分析后再调用 I/O 模块存储相关分析信息。模块涉及的相关类图如图 4-2 所示。

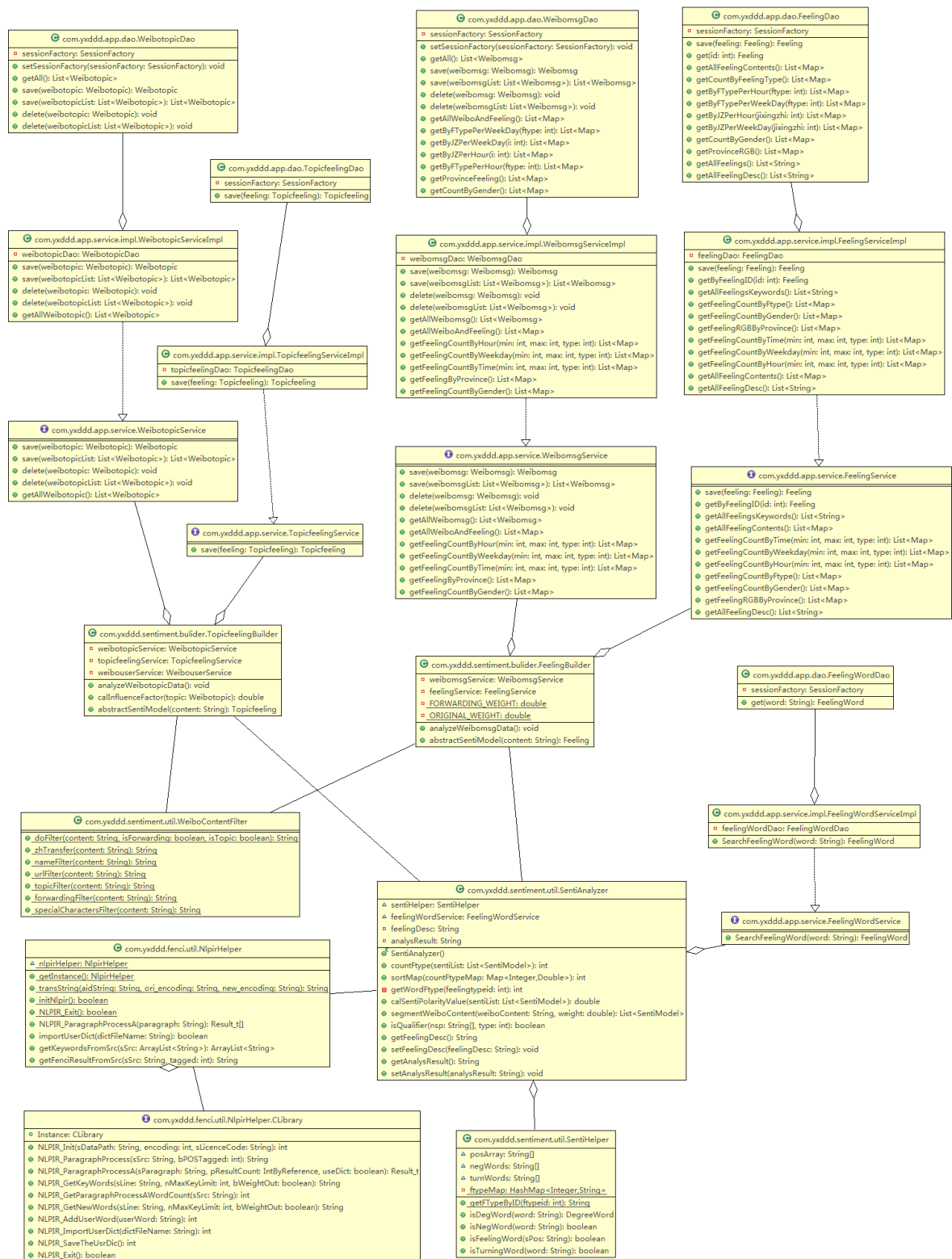


图 4-2 情感分析模块类图

### 4.3.2 噪声过滤模块

噪声过滤模块主要负责在微博内容识别阶段准确地识别出垃圾微博，并将垃圾微博从待情感分析的微博列表中删除。首先需调用 I/O 模块获取微博语料，通过文本处理模块对语料进行预处理后

获取关键词，通过噪声过滤识别出噪声词和垃圾微博后，再调用 I/O 模块存储噪声信息。模块涉及的相关类图如图 4-3 所示。

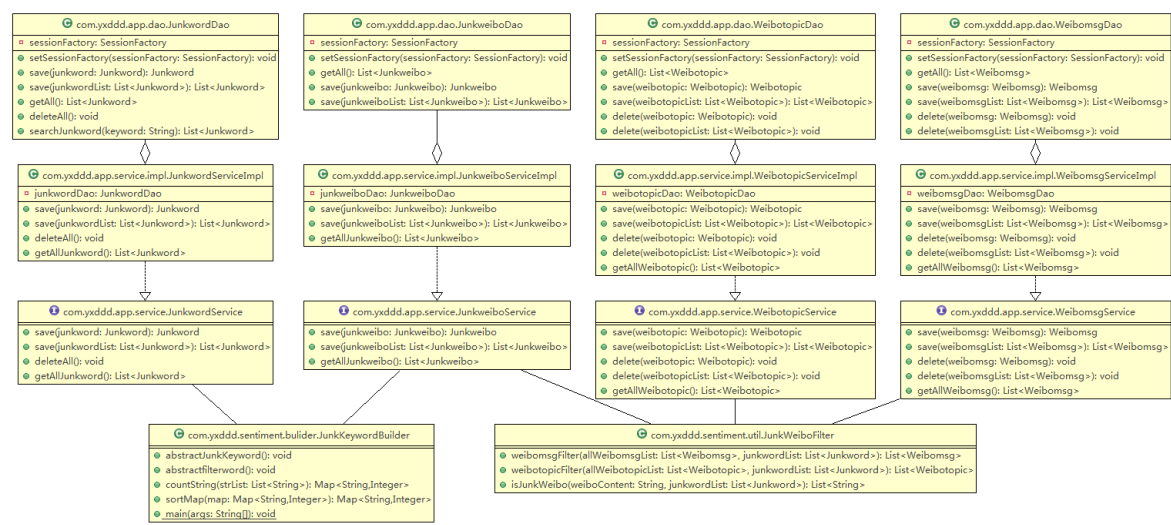


图 4-3 噪声过滤模块类图

### 4.3.3 可视化模块

可视化模块主要负责将微博情感挖掘的处理过程及数据统计以图表的形式显示出来，包括处理过程可视化和处理结果可视化两个部分。

在处理过程可视化中，需调用情感分析模块对前台传过来的文本数据进行情感分析。

在处理结果可视化中，需调用 I/O 模块对情感分析信息进行统计，再将统计数据传至前台界面。

模块用例图如图 4-4 所示：

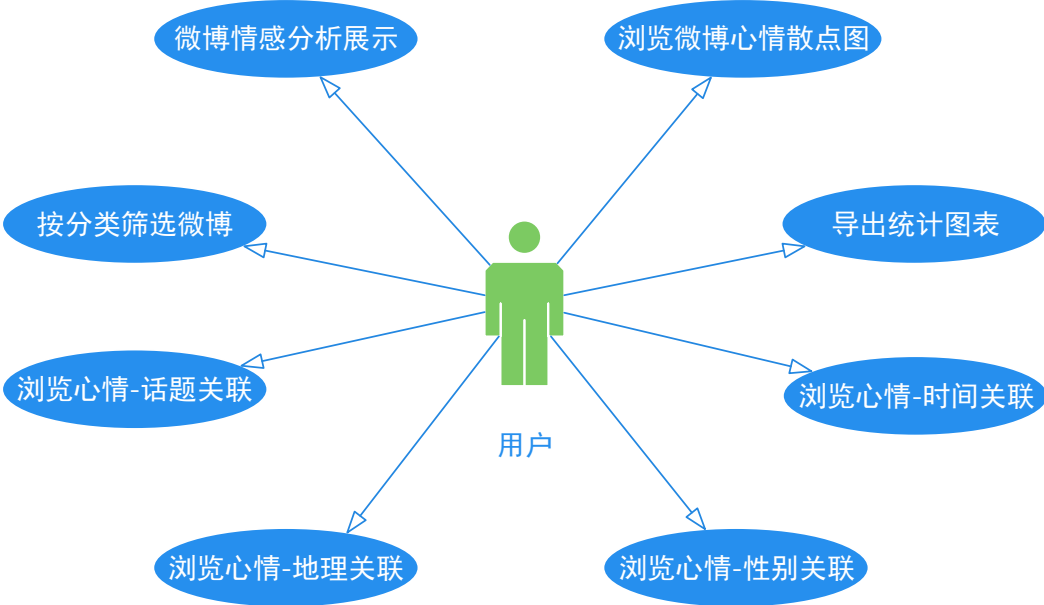


图 4-4 模块用例图

模块涉及的后台相关类图如图 4-5 所示。

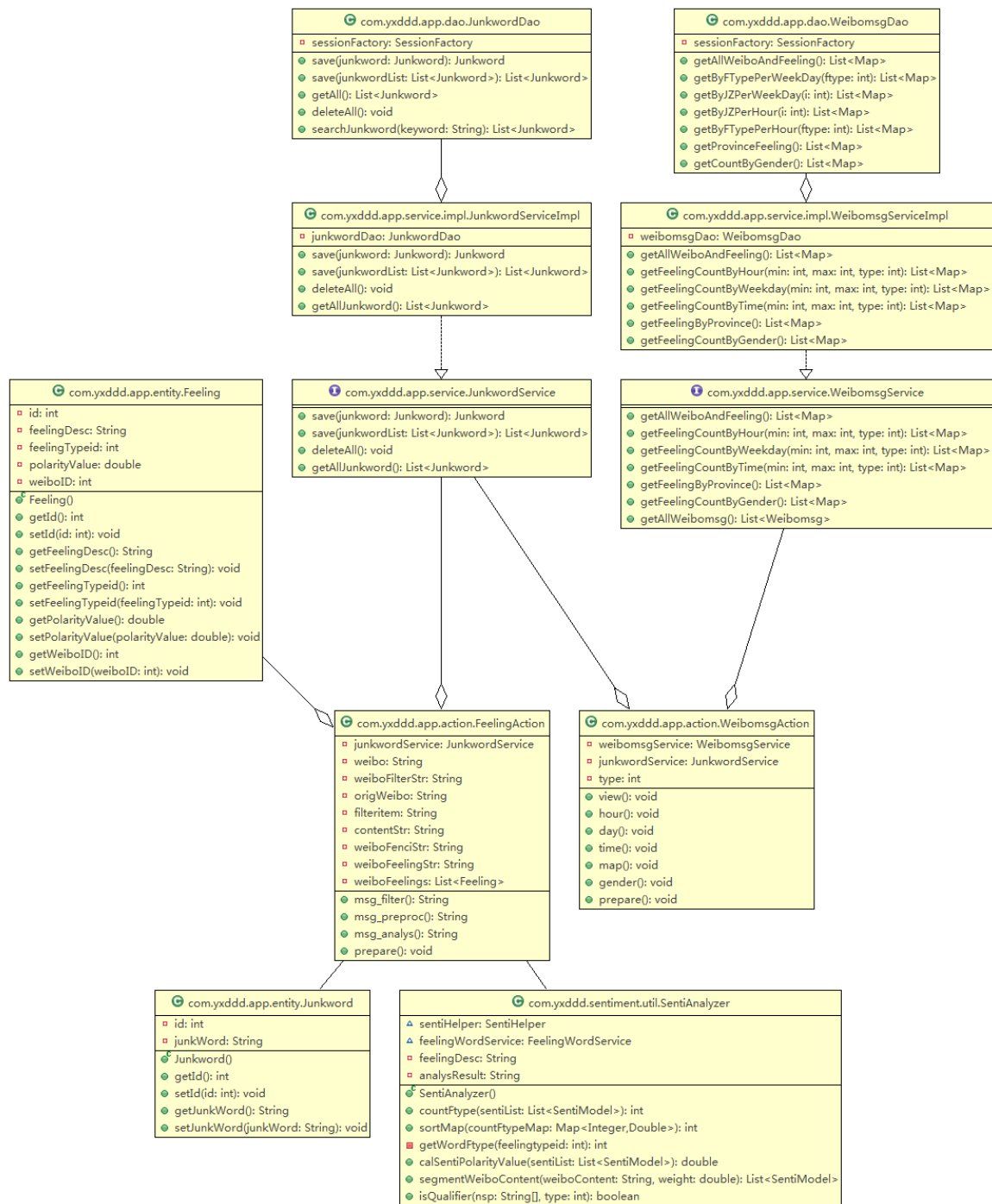



图 4-5 可视化模块类图

## 4.4 关键类说明

### 4.4.1 JunkKeywordBuilder 类说明

JunkKeywordBuilder 的主要作用是遍历数据库中的垃圾微博列表，对每一条垃圾微博进行分词处理，提取出关键词；然后对所有关键词的出现频率排序，频率达到阈值的则存入噪声词库中，词库中已存在的垃圾词不重复录入。



 com.yxddd.sentiment.bulider.JunkKeywordBuilder
<ul style="list-style-type: none"><li>abstractJunkKeyword(): void</li><li>abstractfilterword(): void</li><li>countString(strList: List&lt;String&gt;): Map&lt;String,Integer&gt;</li><li>sortMap(map: Map&lt;String,Integer&gt;): Map&lt;String,Integer&gt;</li></ul>

4.4.1.1 类数据成员

无

4.4.1.2 类成员函数

表 4-13 类成员函数设计表

函数名	abstractJunkKeyword		函数作用范围	Private
类名	JunkKeywordBuilder			
功能概要	遍历垃圾微博，提取出垃圾微博关键字			
记述形式				
参数				
类型	变量名		I/O	说明
无	无		无	无
返回值	类型	无	说明	
	值	无		
详细说明				
此函数为垃圾词的生成器。通过遍历垃圾微博、对垃圾微博进行分词处理后提取垃圾微博的关键词。				
使用注意事项				

abstractJunkKeyword 函数流程图如下：



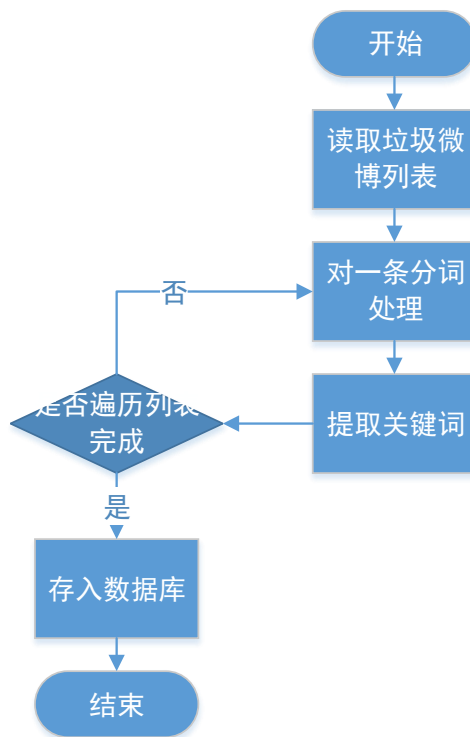


图 4-6 类成员函数流程图

表 4-14 类成员函数设计表

函数名	abstractfilterword		函数作用范围	Private
类名	JunkKeywordBuilder			
功能概要	遍历垃圾微博关键字，提取出频率超过 threshold 的词作为过滤词			
记述形式				
参数				
类型	变量名		I/O	说明
无	无		无	无
返回值	类型	无	说明	
	值	无		
详细说明				
所有关键词存入一个 List 中，统计该 List 中关键词的数量，词频超过 threshold 的存入数据库中，数据库中已存在的关键词不重复录入。				
使用注意事项				

abstractfilterword 函数流程图如下：

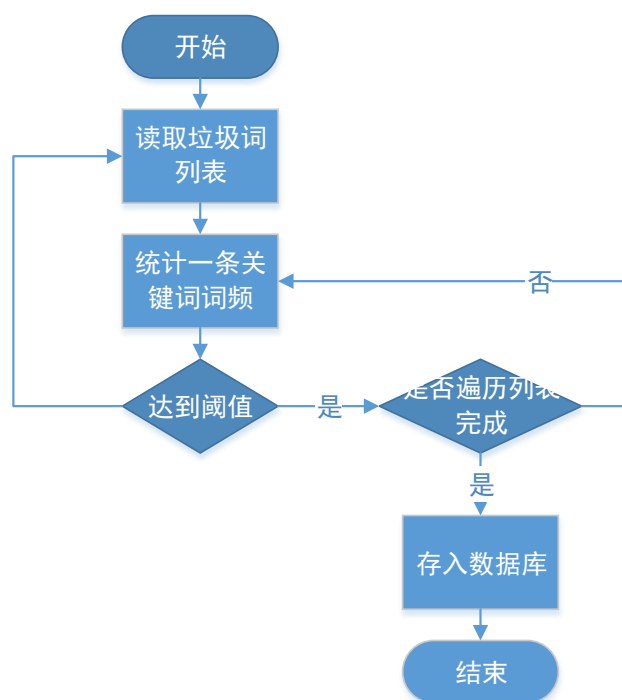


图 4-7 类成员函数流程图

## 4.4.2 FeelingBuilder 类说明

FeelingBuilder 的主要作用是遍历数据库中所有的微博信息，过滤微博中的表情符、网页链接、用户名信息（如@银杏大道东），计算微博的情感极性、强度和情感分类，若带有转发信息的微博另分析转发微博的情感，并以一定的权重计算出最后微博的情感值。

com.yxddd.sentiment.bulider.FeelingBuilder	
weibomsgService: WeibomsgService	
feelingService: FeelingService	
FORWARDING_WEIGHT: double	
ORIGINAL_WEIGHT: double	
analyzeWeibomsgData(): void	
abstractSentiModel(content: String): Feeling	

### 4.4.2.1 类数据成员

变量名	类型	说明
weibomsgService	WeibomsgService	负责 Weibomsg 类的数据库操作
feelingService	FeelingService	负责 Feeling 类的数据库操作
FORWARDING_WEIGHT	private static double	常量，转发微博权重，用以计算含转发微博信息的微博情感

ORIGINAL_WEIGHT	private static double	常量，源微博权重
-----------------	-----------------------	----------

4.4.2.2 类成员函数

表 4-15 类成员函数设计表

函数名	analyzeWeibomsgData		函数作用范围	public
类名	FeelingBuilder			
功能概要	分析微博信息			
记述形式				
参数				
类型	变量名		I/O	获取微博信息列表
无	无		无	无
返回值	类型	无	说明	
	值	无		
详细说明				
获取微博信息列表，对列表进行情感分析，若能匹配任意一种情感则保存情感分析结果，不能则舍去				
使用注意事项				

analyzeWeibomsgData 函数流程图如下：

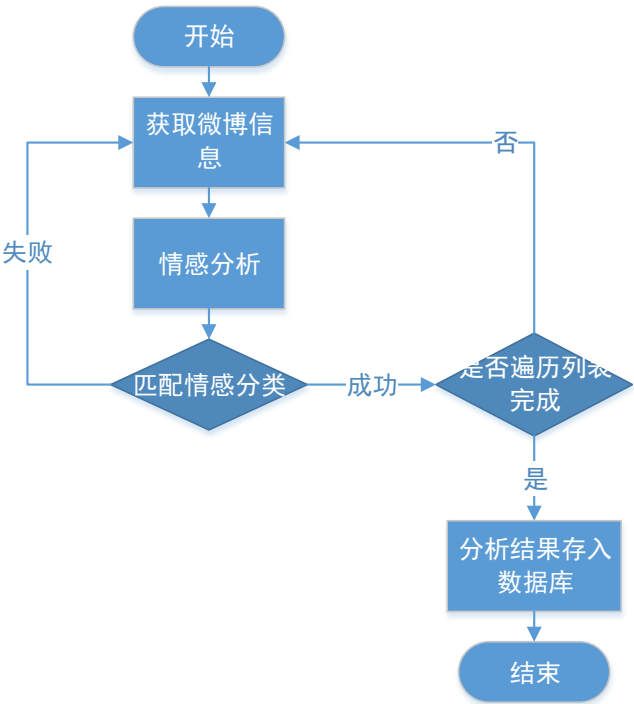


图 4-8 类成员函数流程图

表 4-16 类成员函数设计表

函数名	abstractSentiModel		函数作用范围	public
类名	FeelingBuilder			
功能概要	获取一条微博信息所对应的 Feeling 实体			
记述形式				
参数				
类型	变量名		I/O	获取微博信息列表
String	content		无	无
返回值	类型	Feeling	说明	
	值	feeling	返回该微博文本所属的情感分类的实体。	
详细说明				
过滤微博内容，若微博文本中包含转发内容，拆分文本，对原创及转发内容的情感模型赋予不同权重，获取微博情感极性值、情感分析并返回。				
使用注意事项				

analyzeWeibomsgData 函数流程图如下：

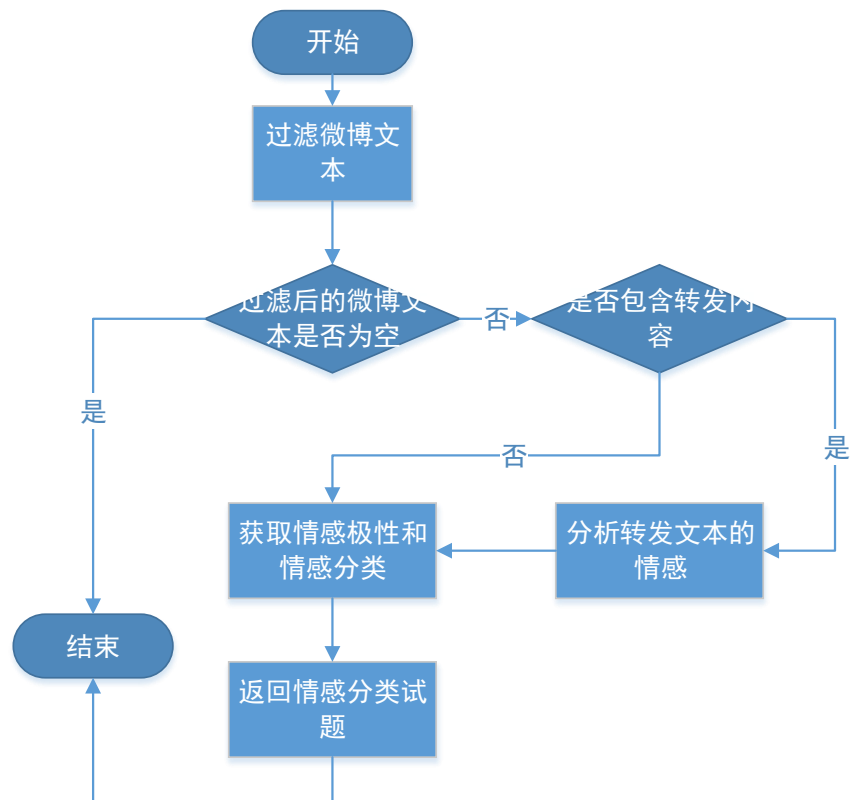






图 4-9 类成员函数流程图

4.4.3TopicfeelingBuilder 类说明

 com.yxddd.sentiment.bulider.TopicfeelingBuilder
<div><div>▪</div>weibotopicService: WeibotopicService</div> <div><div>▪</div>topicfeelingService: TopicfeelingService</div> <div><div>▪</div>weibouserService: WeibouserService</div>
<div><div></div>analyzeWeibotopicData(): void</div> <div><div></div>calInfluenceFactor(topic: Weibotopic): double</div> <div><div></div>abstractSentiModel(content: String): Topicfeeling</div>

4.4.3.1 类数据成员

变量名	类型	说明
weibotopicService	WeibotopicService	负责 Weibotopic 类的数据操作
feelingService	TopicfeelingService	负责 Topicfeeling 类的数据操作
weibouserService	WeibouserService	负责 Weibouser 类的数据操作

4.4.3.2 类成员函数

表 4-17 类成员函数设计表

函数名	analyzeWeibotopicData		函数作用范围	public
类名	TopicfeelingBuilder			
功能概要	分析微博话题信息			
记述形式				
参数				
类型	变量名		I/O	获取微博话题信息列表
无	无		无	无
返回值	类型	无	说明	
	值	无		
详细说明				
获取微博话题信息列表，对列表进行情感分析，若能匹配任意一种情感则保存情感分析结果，不能则舍去				

使用注意事项

analyzeWeibotopicData 函数流程图如下：

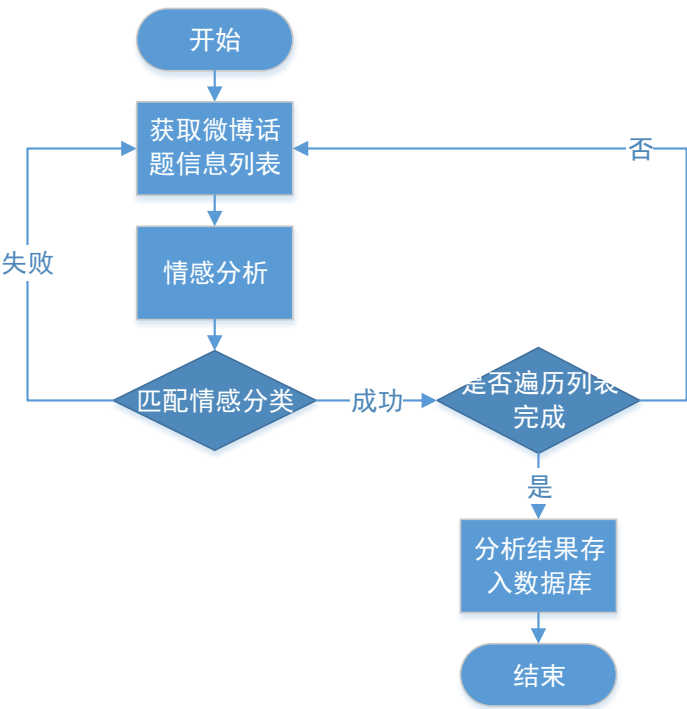


图 4-10 类成员函数流程图

表 4-18 类成员函数设计表

函数名	calInfluenceFactor		函数作用范围	public
类名	TopicfeelingBuilder			
功能概要	计算微博影响力权重			
记述形式				
参数				
类型	变量名		I/O	
Weibotopic	topic		无	无
返回值	类型	double	说明	
	值	ifactor	返回值为该条微博的影响力因素	
详细说明				
<p>根据微博的转发数，评论数，赞数计算该微博的影响力。</p> <p>微博内容影响力公式：<math>MI = \sqrt[3]{MF} + \sqrt{MC} + \sqrt{MP}</math></p> <p>MI 为该条微博内容影响力因数，MF 为该条微博转发数，MC 为该条微博评论数，MP 为该条微博赞数。</p>				

使用注意事项

calInfluenceFactor 函数流程图如下：



图 4-11 类成员函数流程图

表 4-19 类成员函数设计表

函数名	abstractSentiModel	函数作用范围	public
类名	TopicfeelingBuilder		
功能概要	获取一条微博信息所对应的 Topicfeeling 实体		
记述形式			
参数			
类型	变量名		I/O
String	context		无
返回值	类型	Topicfeeling	说明
	值	feeling	微博信息所对应的 Topicfeeling 实体
详细说明			
对微博话题内容（参数）进行分词，获取每个情感词对应的情感模型，根据每个情感词的情感模型，计算此条微博的极性值、情感强度、情感分类，以 Topicfeeling 的实体类型返回。			
使用注意事项			

abstractSentiModel 函数流程图如下：

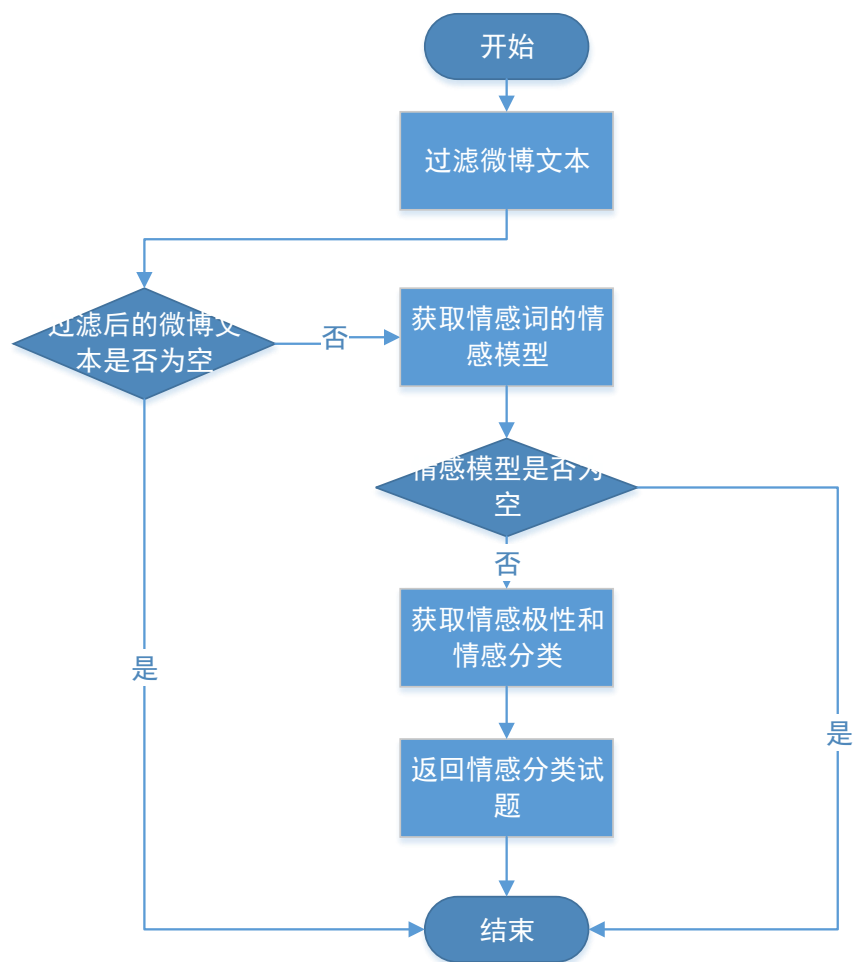


图 4-12 类成员函数流程图

#### 4.4.4 SentiAnalyzer 类说明

com.yxddd.sentiment.util.SentiAnalyzer	
▲ sentiHelper: SentiHelper	
▲ feelingWordService: FeelingWordService	
■ feelingDesc: String	
● SentiAnalyzer()	
● countFtype(sentiList: List<SentiModel>): int	
● sortMap(countFtypeMap: Map<Integer,Double>): int	
● calSentiPolarityValue(sentiList: List<SentiModel>): double	
● segmentWeiboContent(weiboContent: String, weight: double): List<SentiModel>	
● isQualifier(nt: Result_t, type: int): boolean	

##### 4.4.4.1 类数据成员

变量名	类型	说明
-----	----	----



sentiHelper	SentiHelper	负责得到停用词库，程度副词。
feelingWordService	FeelingWordService	负责 FeelingWord 类的数据操作
feelingDesc	String	微博中包含某种情感的情感词

#### 4.4.4.2 类成员函数

表 4-20 类成员函数设计表

函数名	countFtype		函数作用范围	public
类名	SentiAnalyzer			
功能概要	计算微博中各个情感分类的权重, 给出该条微博的情感分类 id			
记述形式				
参数				
类型	变量名		I/O	无
List<SentiModel>	sentiList		无	无
返回值	类型	int	说明	
	值	feelingtypeid	微博情感分类 id	
详细说明				
<p>根据传过来的参数：一条微博中各个情感词的模型，分析各个情感词的分类，得到所有情感词的分类后排序，数量最多的作为该微博的情感分类，返回该情感分类的实体类。</p>				
使用注意事项				

countFtype 函数流程图如下：

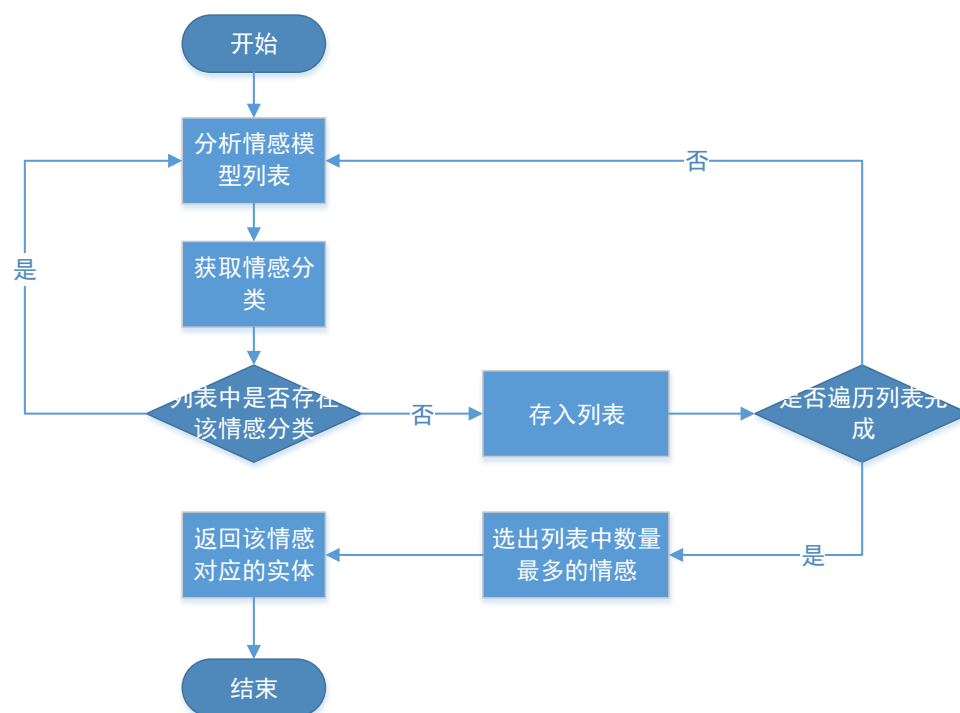


图 4-13 类成员函数流程图

表 4-21 类成员函数设计表

函数名	calSentiPolarityValue		函数作用范围	public
类名	SentiAnalyzer			
功能概要	计算微博极性值			
记述形式				
参数				
类型	变量名		I/O	无
List<SentiModel>	sentiList		无	无
返回值	类型	double	说明	
	值	polarityValue	微博极性值	
详细说明				
遍历情感模型列表，获取情感词极性并累加，返回计算得出的极性值。				
使用注意事项				

calSentiPolarityValue 函数流程图如下：

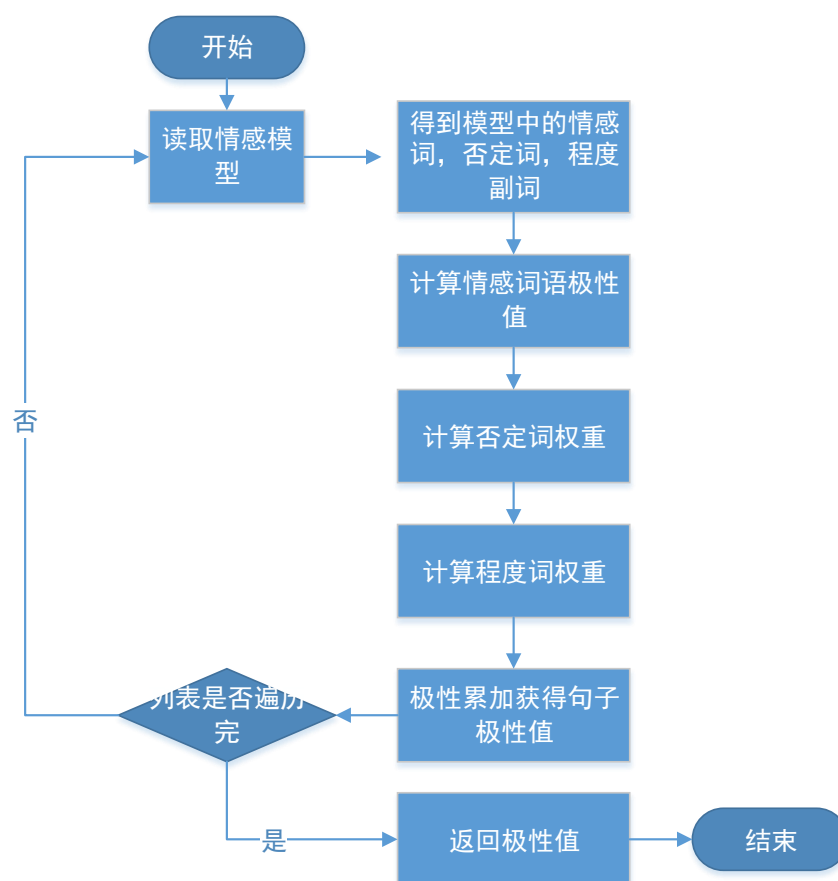


图 4-14 类成员函数流程图

表 4-22 类成员函数设计表

函数名	segmentWeiboContent		函数作用范围	public
类名	SentiAnalyzer			
功能概要	计算微博极性值			
记述形式				
参数				
类型	变量名		I/O	无
String	weiboContent		经过过滤的微博文本	
double	weight			
返回值	类型	List<SentiModel>	说明	
	值	sentiList	情感词模型列表	
详细说明				
遍历微博分词结果实体列表，获取情感词；在每个情感模型中添加程度词列表及否定词列表，判定当前情感词的修饰词位置范围，获取当前情感模型情感词的否定词及程度词，对修饰词进行消歧处理，在当前情感模型中添加情感词的修饰词列表。				
使用注意事项				

segmentWeiboContent 函数流程图如下：

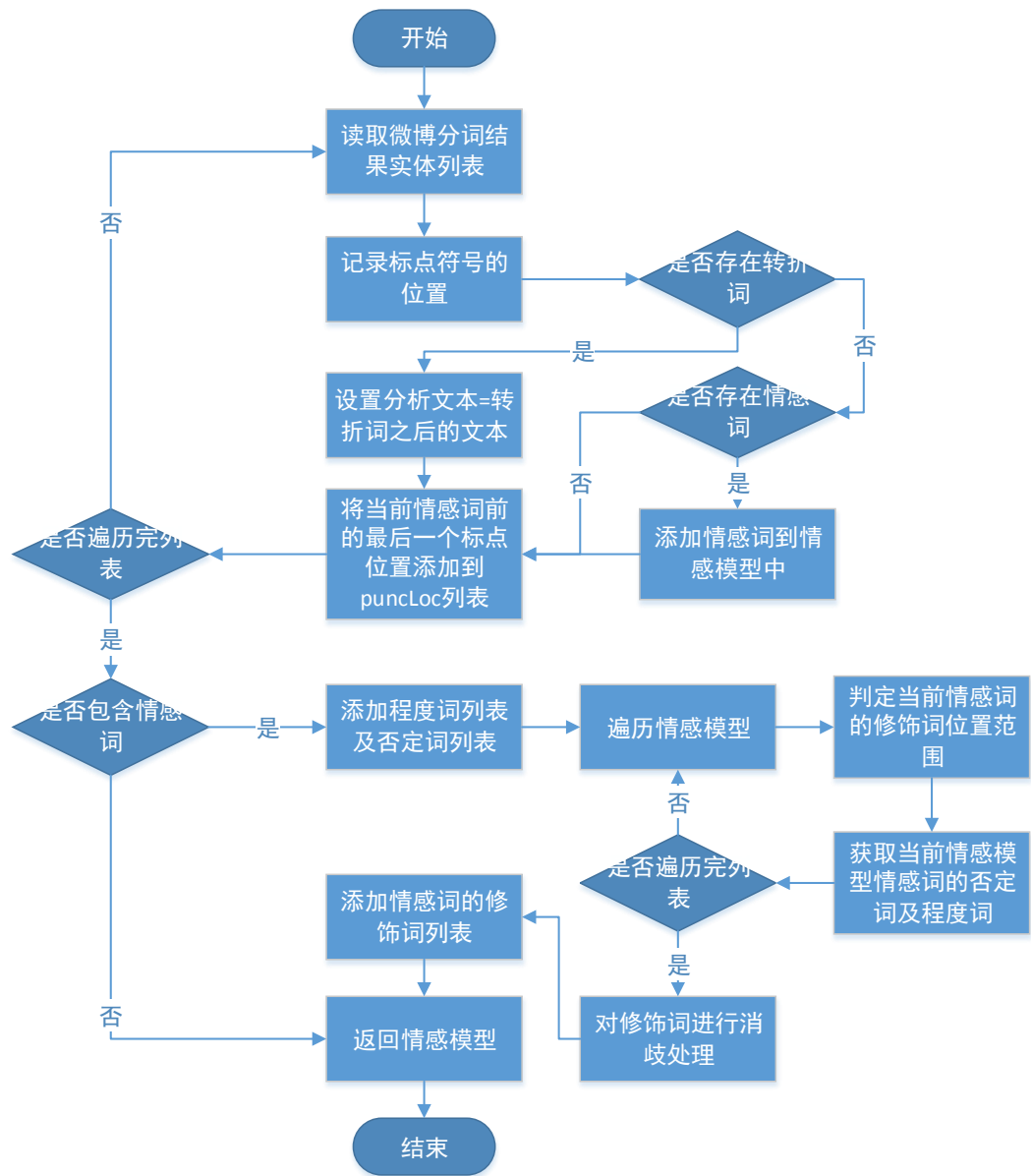


图 4-15 类成员函数流程图

4.4.5WeiboContentFilter 类说明

com.yxddd.sentiment.util.WeiboContentFilter	
doFilter(content: String, isForwarding: boolean, isName: boolean, isTopic: boolean, isSpecial: boolean): String	
nameFilter(content: String): String	
urlFilter(content: String): String	
topicFilter(content: String): String	
forwardingFilter(content: String): String	
specialCharactersFilter(content: String): String	

4.4.5.1 类数据成员

无

4.4.5.2 类成员函数

表 4-23 类成员函数设计表

函数名	doFilter		函数作用范围	public
类名	WeiboContentFilter			
功能概要	总体过滤及繁体转简体			
记述形式				
参数				
类型	变量名		I/O	无
String	content		微博文本	
boolean	isForwarding		是否启用转发内容过滤	
boolean	isName		是否启用用户名过滤	
boolean	isTopic		是否启用话题标签过滤	
boolean	isSpecial		是否启用特殊符号过滤 – 如果使用高级分词则必须过滤，普通分词随意	
返回值	类型	String	说明	
	值	content	过滤后的微博文本	
详细说明				
根据传过来的参数列表，总体过滤及繁体转简体。是否启用转发内容过滤、用户名过滤、话题标签过滤、特殊符号过滤为可选项。				
使用注意事项				

doFilter 函数流程图如下：

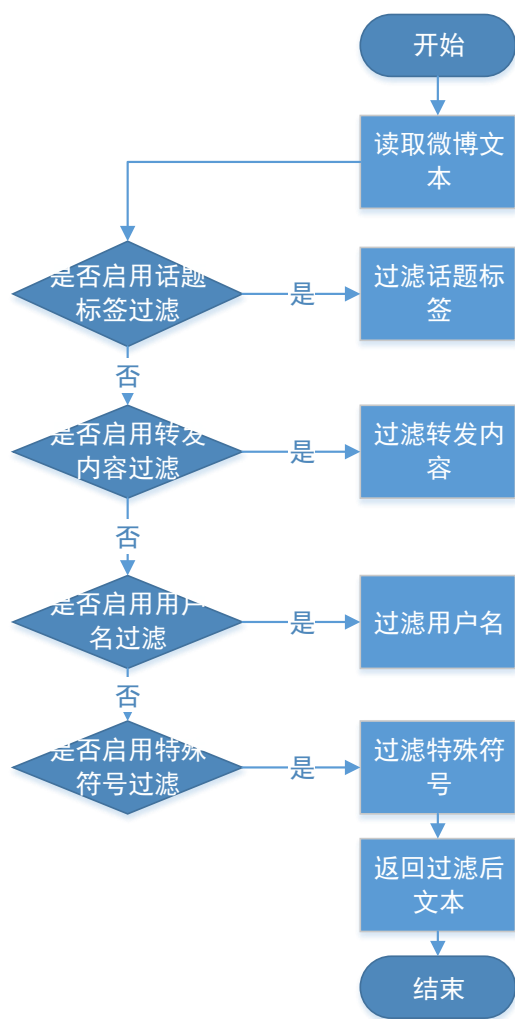


图 4-16 类成员函数流程图

表 4-24 类成员函数设计表

函数名	nameFilter		函数作用范围	public
类名	WeiboContentFilter			
功能概要	过滤微博用户名 @XXX			
记述形式				
参数				
类型	变量名		I/O	无
String	content		微博文本	
返回值	类型	String	说明	
	值	content	过滤用户名后的文本	
详细说明				
根据传过来的微博文本，过滤掉微博用户名 @XXX				
使用注意事项				

nameFilter 函数流程图如下：

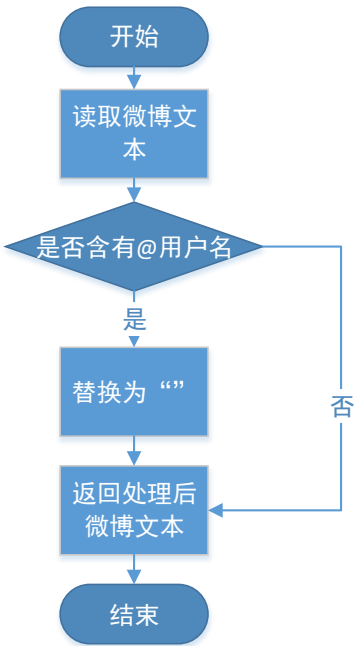






图 4-17 类成员函数流程图

### 4. 4. 6JunkWeiboFilter 类说明

JunkWeiboFilter 的主要作用是遍历数据库中的微博列表，对每一条微博匹配垃圾关键词，若能匹配上则标注这条微博为垃圾微博，不能则为普通微博。

 com.yxddd.sentiment.util.JunkWeiboFilter
 weibomsgFilter(): void
 weibotopicFilter(): void
 isJunkWeibo(weiboContent: String): boolean

#### 4. 4. 6. 1 类数据成员

变量名	类型	说明
junkwordService	JunkwordService	负责 Junkword 类的数据操作
weibotopicService	WeibotopicService	负责 Weibotopic 类的数据操作
weibomsgService	WeibomsgService	负责 Weibomsg 类的数据操作

4.4.6.2 类成员函数

表 4-25 类成员函数设计表

函数名	weibomsgFilter		函数作用范围	public
类名	JunkWeiboFilter			
功能概要	遍历 weibomsg 表，删除包含 junkword 的垃圾微博			
记述形式				
参数				
类型	变量名		I/O	无
List<Weibomsg>	allWeibomsgList		全部微博列表	
List<Junkword>	junkwordList		全部垃圾关键词列表	
返回值	类型	List<Weibomsg>	说明	
	值	junkWeibomsgList	垃圾微博列表，待删除	
详细说明				
遍历 weibotopic 表，删除包含 junkword 的垃圾微博，删除 msgContent 字段为 Null、只包含“转发微博”、createTime 字段为 Null 的垃圾微博				
使用注意事项				

weibomsgFilter 函数流程图如下：

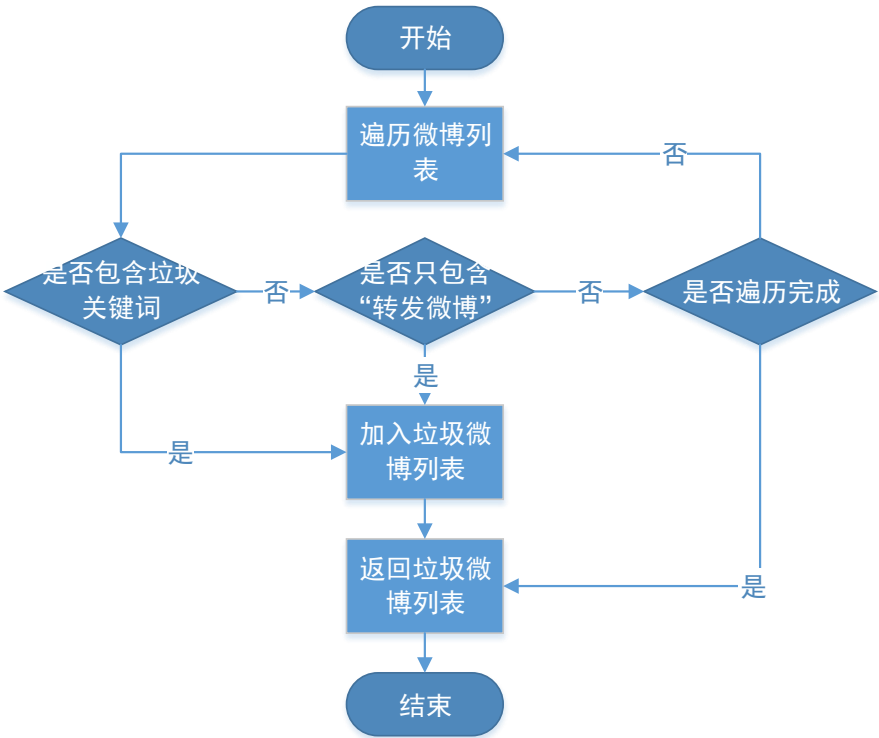


图 4-18 类成员函数流程图



## 4.4.7WeibomsgAction 类说明

WeibomsgAction 的主要作用负责与界面的交互，响应界面的点击、提交表单事件，调用后台相应方法响应服务的请求。

### 4.4.7.1 类数据成员

变量名	类型	说明
response	HttpServletResponse	
type	int	type=1 表示不同的心情分类； type=2 表示极性值
weibomsgService	WeibomsgService	负责 Weibomsg 类的数据操作

### 4.4.7.2 类成员函数

表 4-26 类成员函数设计表

函数名	time		函数作用范围	public
类名	WeibomsgAction			
功能概要	查看一周中每天不同微博心情总量			
记述形式				
参数				
类型	变量名		I/O	无
无	无		无	
返回值	类型	无	说明	
	值	无	无	
详细说明				
响应页面的点击事件，返回每种心情 24 小时和一周的分布数量，以 json 的形式传回前台并展示。				
使用注意事项				

time 函数功能时序图如下：

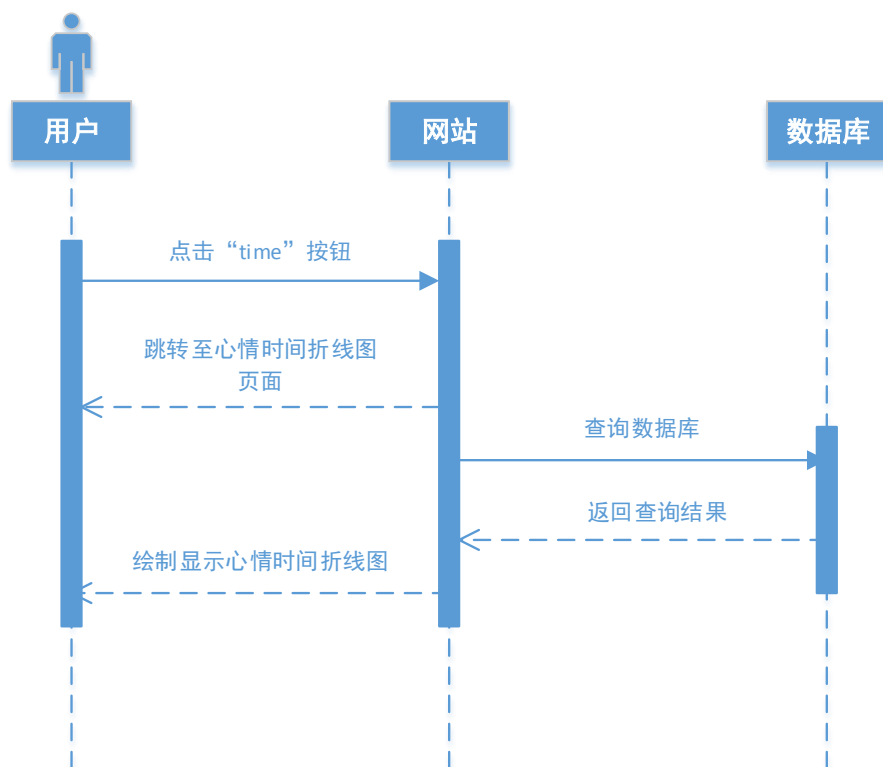


图 4-19 函数功能时序图

表 4-27 类成员函数设计表

函数名	map		函数作用范围	public
类名	WeibomsgAction			
功能概要	查看不同性别心情总量			
记述形式				
参数				
类型	变量名		I/O	无
无	无		无	
返回值	类型	无	说明	
	值	无	无	
详细说明				
响应页面的点击事件，返回不同地理位置的心情的总量，以 json 的形式传回前台并展示。				
使用注意事项				

map 函数功能时序图如下：

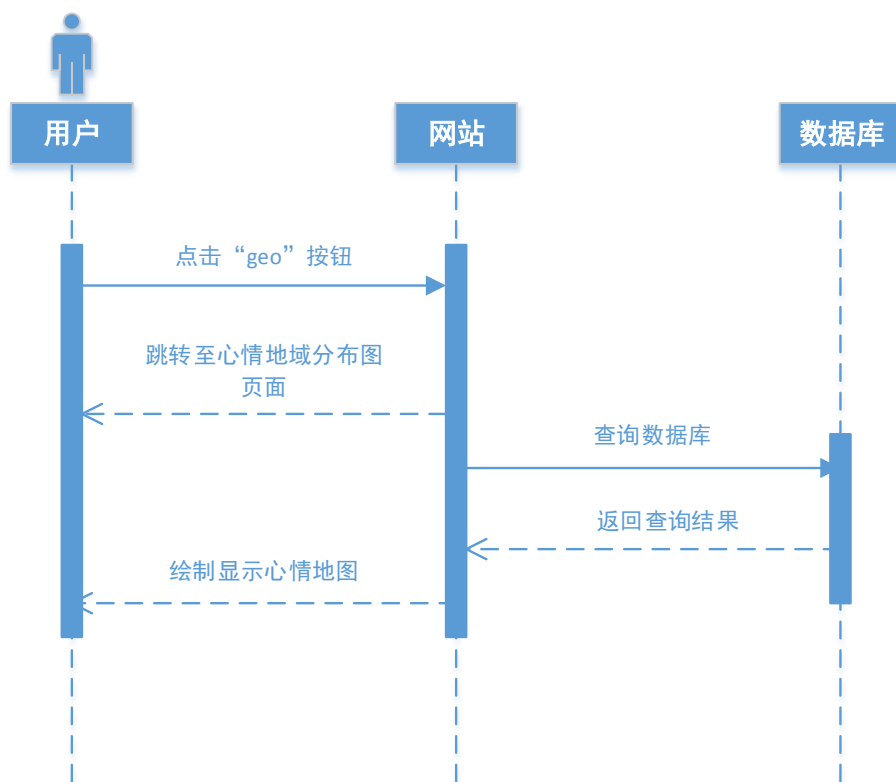


图 4-20 函数功能时序图

## 4.4.8 ViewScatter 类说明

### 4.4.8.1 类数据成员

略

### 4.4.8.2 类成员函数

表 4-28 类成员函数设计表

函数名	processData		函数作用范围	Private
类名	ViewScatter			
功能概要	将传过来的 JSON 转化成点数据，并显示。			
记述形式				
参数				
类型	变量名		I/O	说明
var	dt		无	无
返回值	类型	var	说明	

	值	data	
详细说明			
将传过来的 JSON 转化成点数据，储存点的大小，初始化圆形的位置。			
使用注意事项			

processData 函数流程图如下：

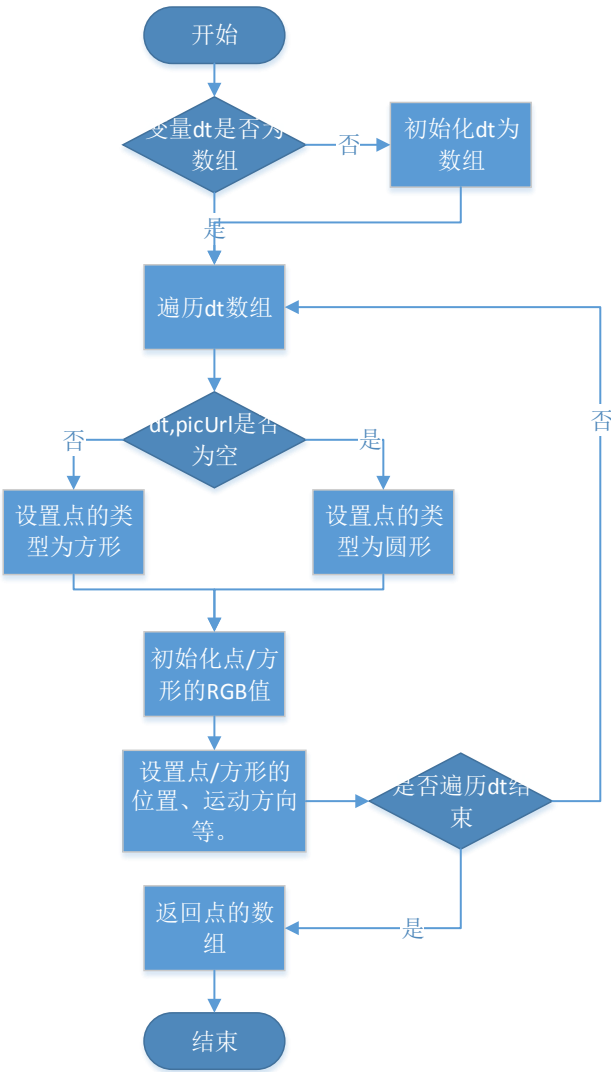


图 4-21 类成员函数流程图

表 4-29 类成员函数设计表

函数名	render	函数作用范围	Private
类名	ViewScatter		
功能概要	绘制图像，每 40 毫秒刷新一次		
记述形式			

参数				
类型	变量名		I/O	说明
无	无		无	无
返回值	类型	无	说明	
	值	无		
详细说明				
绘制图像，每 40 毫秒刷新一次				
使用注意事项				

render 函数流程图如下：

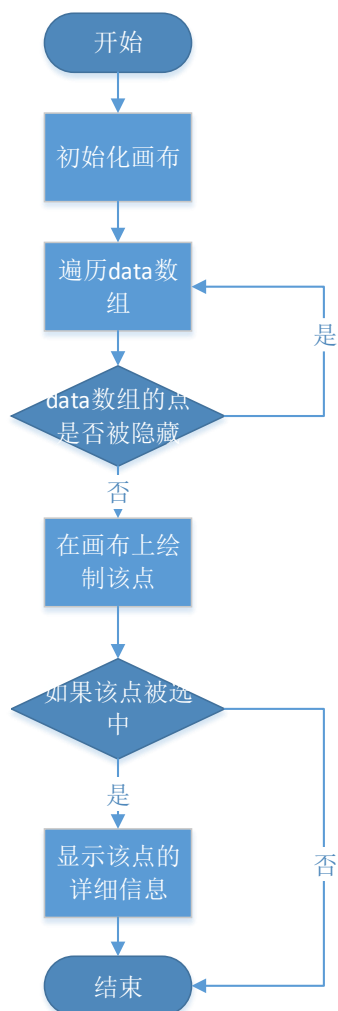


图 4-22 类成员函数流程图

# 5. 接口设计

本系统的整体模型设计采用 B/S 模式，基于 SSH 框架的系统体系结构实现的，客户端通过浏览器访问服务器提供的 Web 页面，与后台本体库交互。

## 5.1 接口设计概述

接口是提供给其他模块或者系统使用的一种约定或者规范。因此接口必须要保证足够的稳定性和易用性。

下表描述了设计接口的基本要求：

表 5-1 接口设计基本要求表

需求名称	需求描述
稳定性	接口必须相对稳定,否则将导致接口的使用者和提供者为了适应新接口而不断修改接口的实现,可能重复进行无用功,严重时影响整个软件开发进度。
易用性	采用面向对象思想。
规范性	主要是接口设计的代码规范,这是最基本的要求。同时考虑 C 接口命名污染的问题。一般 C 接口都会在接口前加上公司或者模块的标识。
可移植性	对于需要在多平台实现的接口需要考虑接口本身的可移植性,因此最少使用对于系统依赖的类型作为接口的参数类型或者返回值类型。
鲁棒性	接口需要有适度的鲁棒性,主要是指能够在多种情况下接口都能实现统一的效果,不会随着调用者传入的参数变化而导致接口的输出出现违背接口语义的情况出现。
安全性	接口定义时需要严格限制参数的读写权限,如果只能是只读的参数一定要设置成 const,以免出现非法使用。
兼容性	这是接口扩充的原则,必须保证同一个接口实现后向兼容前一版本的使用。扩充的同类接口也能兼容老接口的实现。

## 5.2 用户接口

本系统提供可视化的操作方式，不提供命令控制语句进行输入控制，从而用户只需要使用鼠标进行命令操作。

用户主要通过窗体、控件等可视化元素进行交互。

考虑到本系统的特性，用户界面应符合以下设计规范：

表 5-2 用户界面设计规范表

要求	说明
友好性	界面直观、对用户透明，效果炫丽

交互性	支持各种图表操作：包括图表拖拽、导入/导出、放大/缩小、关键实体强调等。
跨平台	支持在移动终端设备上数据进行可视化

## 5.3 外部接口

### 5.3.1 软件接口

服务器程序可使用 Spring 提供的对 MySQL SERVER 的接口，进行对数据库的所有访问。服务器程序上可使用 MySQL SERVER 的数据库对重要数据进行的备份，防止数据被破坏而无法恢复。服务器端采用 Java 语言来编写程序，通过 JDBC 驱动来访问数据库。

### 5.3.2 硬件接口

在输入方面，对于键盘、鼠标的输入，可用 jsp 的标准输入/输出，对输入进行处理。

## 5.4 内部接口

设计内部接口时，各模块之间主要采用函数调用，函数传递，返回值的方式进行信息传递。

## 6. 运行设计

### 6.1 运行模块组合

客户机程序在有输入时启动接收数据模块，通过各模块之间的调用，读入并对输入进行格式化。在接收数据模块得到充分的数据时，将调用网络传输模块，将数据通过网络送到服务器，并等待接收服务器返回的信息。接收到返回信息后随即调用数据输出模块，对信息进行处理，产生相应的输出。

服务器程序的接收网络数据模块必须始终处于活动状态。接收到数据后，调用数据处理/查询模块对数据库进行访问，完成后调用网络发送模块，将信息返回客户机。

### 6.2 运行控制

运行控制将严格按照各模块间函数调用关系来实现。在各事务中心模块中，需对运行控制进行正确的判断，选择正确的运行控制路径。

在网络传输方面，客户机在发送数据后，将等待服务器的确认收到信号，收到后，再次等待服务器发送回答数据，然后对数据进行确认。服务器在接到数据后发送确认信号，在对数据处理，访问数据库后，将返回信息送回客户机，并等待确认。

### 6.3 运行时间

在软体的需求分析中，对运行时间的要求为必须对做出的操作有较快的反应。网络硬件对运行时间有最大的影响，当网络负载量大时，对操作反应将受到很大的影响。其次是服务器的性能，这将影响对数据库访问时间即操作时间的长短，影响加大客户机操作的等待时间，所以必须使用高性能的服务器。



## 7. 系统出错处理设计

### 7.1 出错信息

在错误发生时，给出出错的原因。

### 7.2 补救措施

系统遭到恶意攻击，或是中毒以后，导致系统内部数据紊乱，用户信息失真，因此，应定期对系统数据进行自动备份，以便数据丢失时，能第一时间恢复，减少损失；另外，要定期对整个系统进行纸制的备份，以免出现自然灾害而导致所有数据的丢失，自动制定自动还原点；

在网络传输方面，可考虑建立一条成本较低的后备网络，以保证当主网络断路时的据之通信。

在硬件方面要选择较可靠，稳定的服务器机种，保证系统运行时的可靠性。

### 7.3 系统维护设计

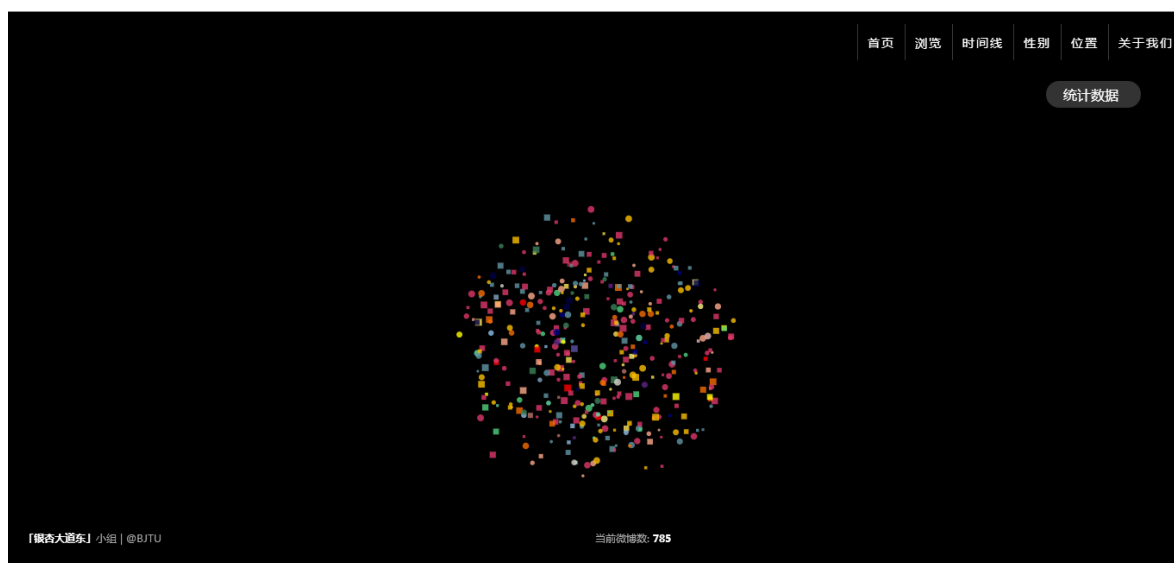
采用模块化的设计，方便维护。

## 8. 运行效果展示

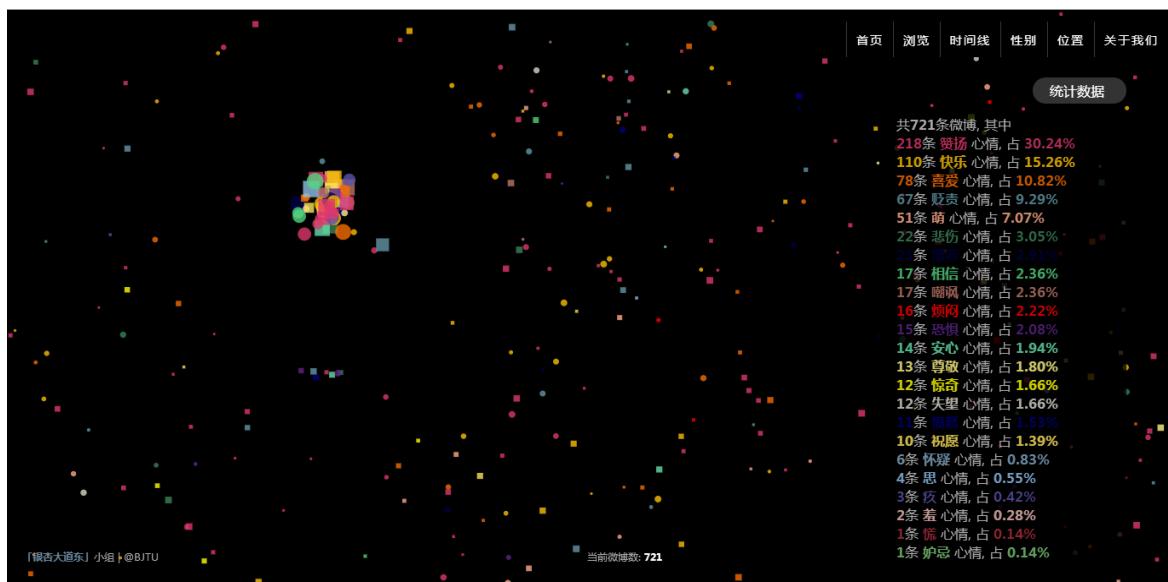
- ♥ 首页：“今天，你是什么颜色”，项目说明。



- ♥ 心情浏览：每个点/方形代表一条微博信息，不同的颜色代表不同的心情。随机显示 700-800 条微博信息。方形代表微博含有图片信息。



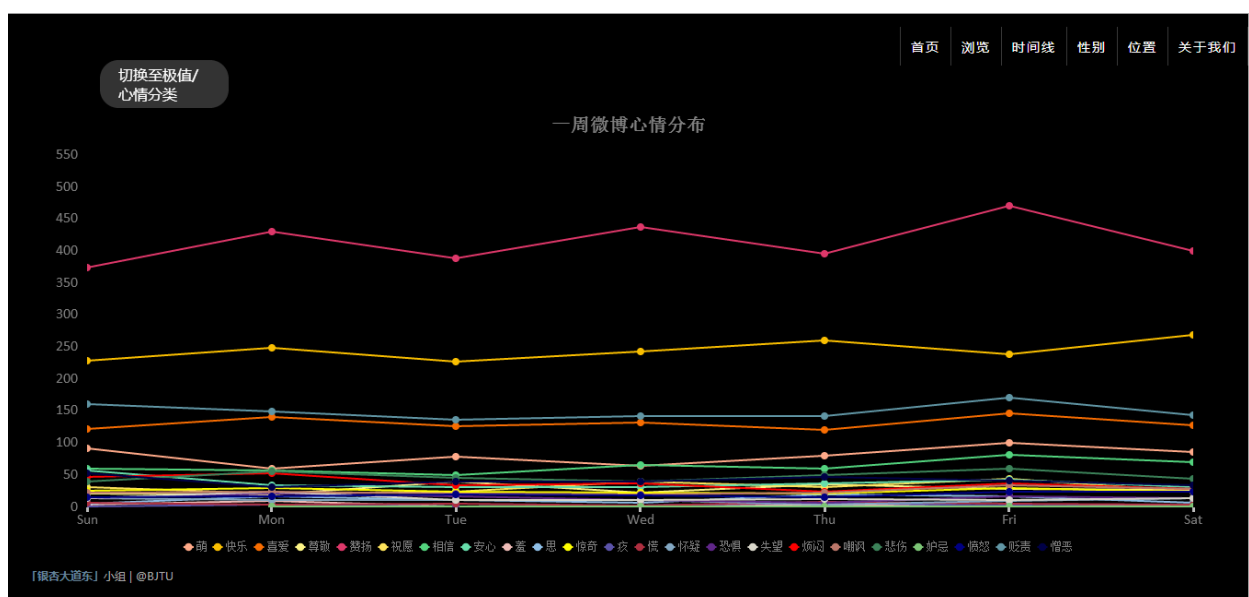
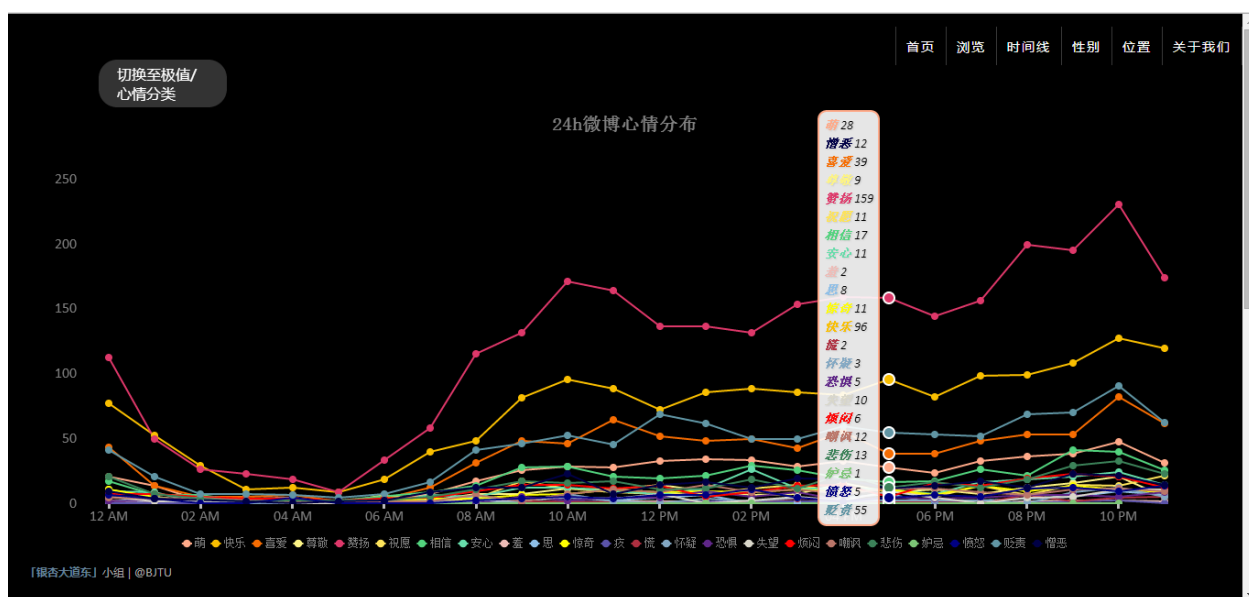
- ♥ 鼠标点击右上方的统计数据，即可显示当前散点图中的微博总数，各心情类型微博的数量及所占百分比。鼠标停留在某一位置，则周围的浮点会向鼠标聚集。



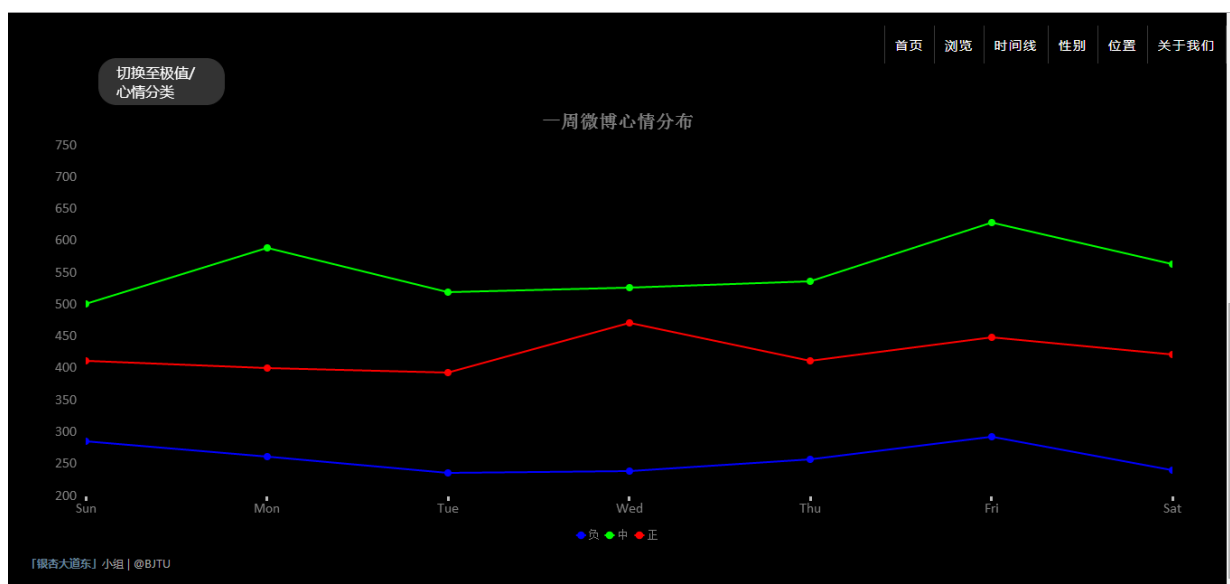
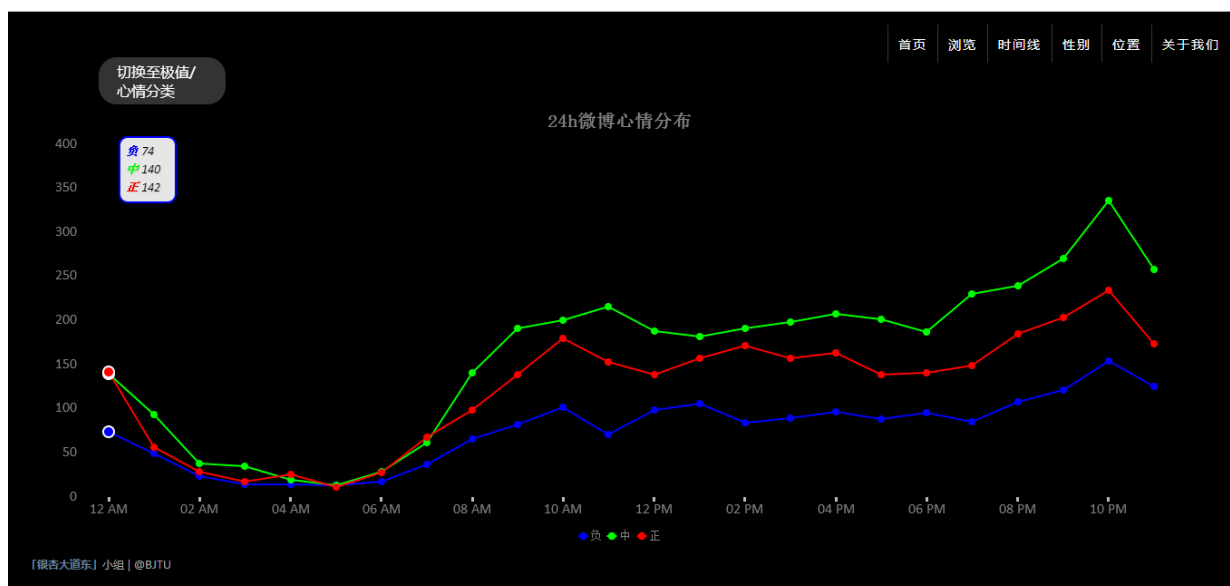
- ♥ 点击某个点/方形，即可显示该条微博的具体信息，由微博心情分类，微博内容，微博发布人昵称，头像和个人主页链接组成。



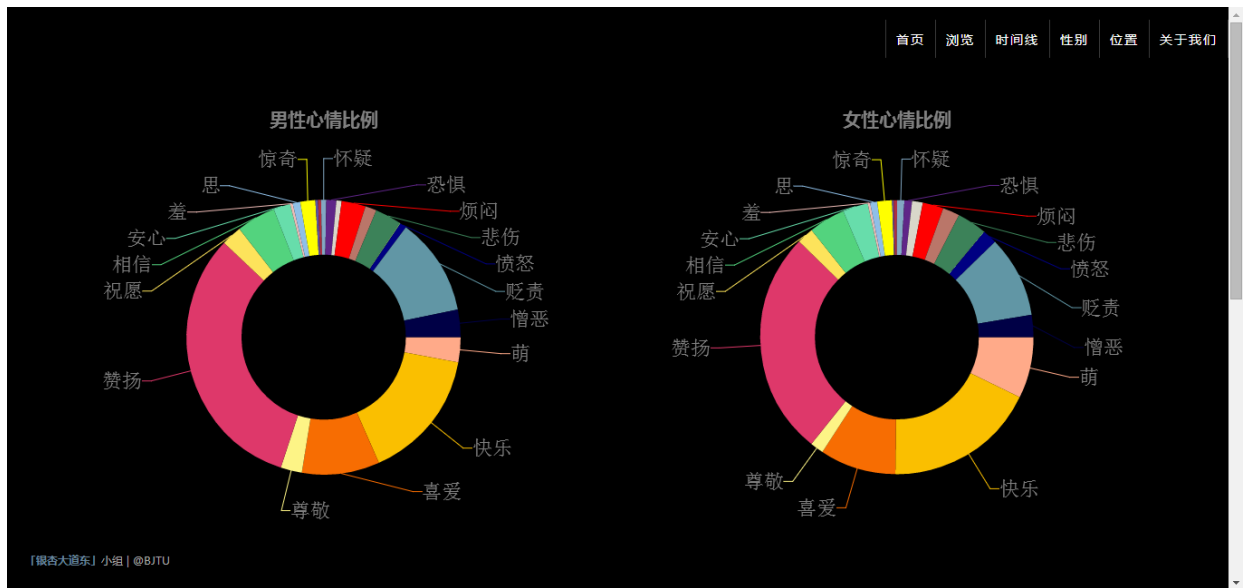
- ♥ 心情类型-时间分布曲线：不同颜色的曲线代表不同心情微博的数量，可以看出同一时刻不同心情微博的数量。鼠标停留在任一时刻可显示所有心情类型微博的具体数值（共分 23 种心情）。共有两种时间显示方式，一种为 24h 心情分布，由 12: 00AM-12: 00PM，以 1 小时为一个时间间隔单位。一种为一周心情分布，从周日到周一，以 1 天为一个时间间隔单位。



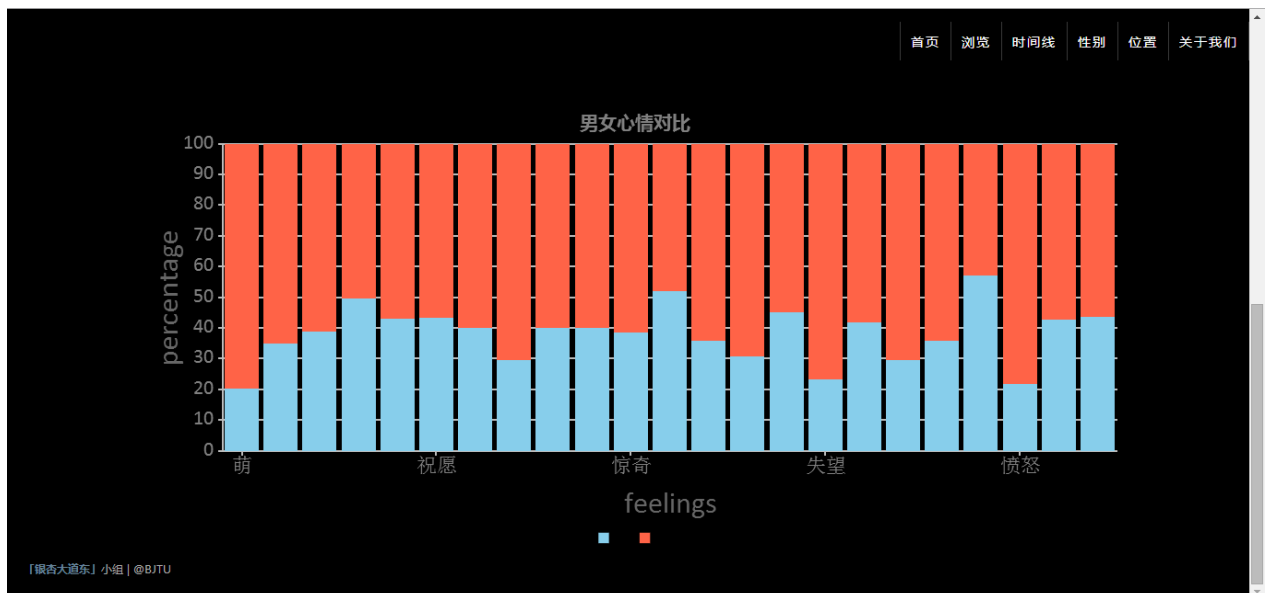
♥ 情感倾向性-时间分布曲线: 显示微博情感倾向性 (正面、中性、负面) 随时间的变化曲线。可以看出同一时刻不同情感倾向性微博的数量。鼠标停留在任一时刻可显示所有情感倾向性微博的具体数值。



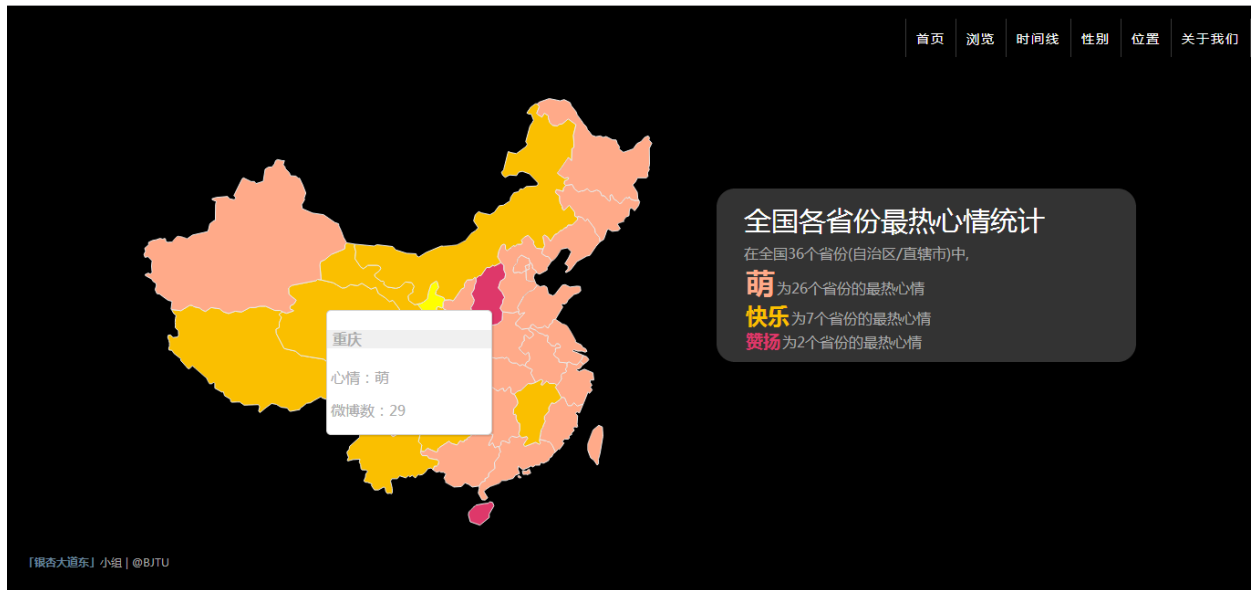
♥ 心情类型-性别分布比例：显示某一性别中，各个心情类型微博所占比例的饼状图，鼠标停留到某一心情区块，即可显示该心情微博数量的悬浮框。



- ♥ 心情类型-男女分布对比：显示各个心情中，男女性别占该心情的百分比。鼠标移至某一区块，即可显示在该心情中男性/女性发布的微博数量。



- ♥ 心情-地理分布：显示全国各省份中最热门的心情，鼠标停留在某一省份，即会出现悬浮框，显示该省份的最热心情及该心情的微博数量。各省颜色为该省最热心情的代表色。


























♥ 情感分析流程展示：显示了微博信息从收集、去噪、简单语义分析、情感分析到可视化的整个流程。下图以“#服创大赛#交稿啦！~~好开心，必胜客走起！！~~@银杏大道东，我在这里:北京交通大学 <http://t.cn/z8A4XyY>”为例，展示了整个流程。



# 9. 附录

## A：心情-颜色对照表

分类	颜色	16 进制值	RGB 值
萌		#FFAA89	255, 170, 137
快乐		#FABF00	250, 191, 0
喜爱		#F76D02	247, 109, 2
尊敬		#FDF486	253, 244, 134
赞扬		#4DCD66	77, 205, 112
祝愿		#FDE35B	253, 227, 91
相信		#53D37E	83, 211, 126
安心		#67DDAB	103, 221, 171
羞		#EEBCB7	238, 188, 183
思		#8FBFE7	143, 191, 231
惊奇		#FFFF00	255, 255, 0
疚		#5B53A9	91, 83, 169
慌		#AE2F40	174, 47, 64
怀疑		#83A8C3	131, 168, 195
恐惧		#5F2688	95, 38, 136
失望		#D9D7CA	217, 215, 202
烦闷		#FF0000	255, 0, 0
嘲讽		#BA7669	186, 118, 105
悲伤		#3C8259	60, 130, 89
妒忌		#7CC576	124, 197, 118
愤怒		#000082	0, 0, 130
贬责		#6196A5	97, 150, 165
憎恶		#000046	0, 0, 70



B：情感分类表

编号	情感大类	情感类	例词
1	乐	萌(PM)	萌萌哒、么么哒、有爱、萌化
2	乐	快乐(PA)	喜悦、欢喜、笑咪咪、欢天喜地
3	乐	喜爱(PB)	倾慕、宝贝、一见钟情、爱不释手
4	好	尊敬(PD)	恭敬、敬爱、毕恭毕敬、肃然起敬
5	好	赞扬(PH)	英俊、优秀、通情达理、实事求是
6	好	祝愿(PK)	渴望、保佑、福寿绵长、万寿无疆
7	好	相信(PG)	信任、信赖、可靠、毋庸置疑
8	好	安心(PE)	踏实、宽心、定心丸、问心无愧
9	惧	羞(NG)	害羞、害臊、面红耳赤、无地自容
10	哀	思(PF)	思念、相思、牵肠挂肚、朝思暮想
11	惊	惊奇(PC)	奇怪、奇迹、大吃一惊、瞠目结舌
12	哀	疚(NH)	内疚、忏悔、过意不去、问心有愧
13	惧	慌(NI)	慌张、心慌、不知所措、手忙脚乱
14	恶	怀疑(NL)	多心、生疑、将信将疑、疑神疑鬼
15	惧	恐惧(NC)	胆怯、害怕、担惊受怕、胆颤心惊
16	哀	失望(NJ)	憾事、绝望、灰心丧气、心灰意冷
17	恶	烦闷(NE)	憋闷、烦躁、心烦意乱、自寻烦恼
18	恶	嘲讽(NM)	蛇精病、逗逼、图样图森破、深井冰
19	哀	悲伤(NB)	忧伤、悲苦、心如刀割、悲痛欲绝
20	恶	妒忌(NK)	眼红、吃醋、醋坛子、嫉贤妒能
21	怒	愤怒(NA)	气愤、恼火、大发雷霆、七窍生烟
22	恶	贬责(NN)	呆板、虚荣、杂乱无章、心狠手辣
23	恶	憎恶(ND)	反感、可耻、恨之入骨、深恶痛绝

C： 团队组成与角色分工

姓名	角色	分工
周婕	组长	数据获取、数据去噪、建立噪声词库
李慧敏	组员	建立情感词典、情感分析、数据转换
王莹	组员	UI 设计与实现、数据可视化
陈一偲	组员	中文分词、情感分析