硕士学位论文

**基于情感分析的视频弹幕系统的设计与实现**

学位申请人：樊影楠

指导教师：饶元 副教授

类别（领域）：工程硕士（软件工程）

2018年03月

**Design and Implementation of Video Barrage System Based on Sentimental Analysis**

A thesis submitted to

Xi’an Jiaotong University

in partial fulfillment of the requirements

for the degree of

Master of Engineering

By

Yingnan Fan

Supervisor: Assoc. Prof. Yuan Rao

(Software Engineering)

March 2018

**论文题目：基于情感分析的视频弹幕系统的设计与实现**

**类别（领域）：工程硕士（软件工程）**

**学位申请人：樊影楠**

**指导教师：饶元 副教授**

摘 要

摘要应采用三段式结构，即只包括三个自然段。注意：摘要中不能出现图、表、数学公式、化学结构式、非公知公用的符号和术语等。摘要第一段指出从事这项研究工作的目的和意义，实际上是“绪论”的高度浓缩。篇幅建议：该段篇幅固定为四行。摘要第一段需要概括绪论前三节的内容，但是不需要概括绪论第四节的内容。

摘要第二段指出完成的工作，即作者独立进行的研究工作及相应成果的概括性叙述，实际上是第2章至倒数第2章的高度浓缩。篇幅建议：该段篇幅至少为十行，最好为十一行，至多为十二行。摘要第二段应首先概括第2章的内容，建议着墨不宜太多，毕竟第2章只是概述已有的技术或理论，篇幅建议为两行。接下来，摘要第二段应分别概括第3章至第5章的内容。通常情况下，第3章论述分析，第4章论述设计，第5章论述实现（通常也包括测试、验证、运行效果等）。摘要第二段概括第3章至第5章时，篇幅建议均不少于三行。在部分学位申请人眼中，第3章至第5章的重要性确有差异。对于重要性较弱的某一章，摘要第二段对其进行概括时，篇幅建议为两行。千万不能僵化地理解上述篇幅建议！虽然工程硕士学位论文的大致结构相似，但是每个学位申请人毕竟有自己的撰写习惯，相应地各章篇幅必然存在一定的波动。在这种情形下，摘要第二段在概括各章时，相应的篇幅应尽量匹配各章的实际篇幅。

摘要第三段指出获得的主要结论，实际上是“结论与展望”的高度浓缩。篇幅建议：该段篇幅固定为三行。注意：摘要第三段应着重概括结论节的内容；尤其当结论节的内容非常饱满时，强烈建议摘要第三段仅概括结论节的内容。

**关 键 词：**视频弹幕；情感分析；词性情感特征

**论文类型：**应用研究

**Title:** **Design and Implementation of Video Barrage System Based on Sentimental Analysis**

**Professional Fields: Software Engineering**

**Applicant: Yingnan Fan**

**Supervisor: Assoc. Prof. Yuan Rao**

ABSTRACT

First paragraph.

Second paragraph.

Third paragraph.

英文摘要撰写要求如下：

1）用词准确，符合语法；

2）关键词按相应专业的标准术语写出，尽量从《英语主题词表》中摘选；

3）如果论文的主体工作得到了有关基金资助，应用英文在摘要页的页脚处标注：本研究得到某某基金（编号： ）资助；

中文摘要和英文摘要均不要求学位申请人及其指导教师签字。

**KEY WORDS**: Video barrage;Sentiment Analysis;Speech emotion feature

**TYPE OF THESIS**: Application Research

目 录

1 绪论 8

1.1 研究背景与意义 8

1.2 国内外研究现状 8

1.3 论文的主要研究内容 10

1.4 论文的组织结构 10

2 相关理论及技术 12

2.1 文本情感分析常用技术 12

2.1.1 网络爬虫 12

2.1.2常见分词技术 13

2.1.3 词性标注 14

2.2 基于机器学习的情感分析方法 14

2.2.1文本表示模型 14

2.2.2 文本特征提取 15

2.2.3 权重计算 15

2.2.4 分类模型 16

2.3 本章小结 17

3 基于SVM的弹幕文本情感分析算法研究 18

3.1 问题描述 18

3.1.1弹幕文本的特点 18

3.1.2 方法概述 19

3.2 算法描述 20

3.2.1 视频弹幕情感词典库的构建 21

3.2.2弹幕文本预处理 21

3.2.3 情感特征提取 22

3.2.4基于SVM的结合高维混合特征弹幕文本情感倾向性分析算法 24

3.2.5结合客观弹幕文本的情感分类模型 25

3.3 实验结果与分析 28

3.3.1 实验性能指标 28

3.3.2 实验设计与结果分析 29

3.4 本章小结 32

4 基于情感分析的视频弹幕系统需求分析 34

4.1基于情感分析的视频弹幕系统需求描述 34

4.2基于情感分析的视频弹幕系统的分析模型 34

4.2.1基于情感分析的视频弹幕系统的功能模型 34

4.2.2基于情感分析的视频弹幕系统的静态模型 39

4.2.3基于情感分析的视频弹幕系统的行为模型 40

4.3基于情感分析的视频弹幕系统的非功能性需求 45

4.4 本章小结 46

5基于情感分析的视频弹幕系统的设计 47

5.1基于情感分析的视频弹幕系统的架构设计 47

5.2基于情感分析的视频弹幕系统的概要设计 48

5.2.1 功能模块设计 48

5.2.2 类的设计 49

5.3基于情感分析的视频弹幕系统的详细设计 51

5.3.1 视频查询模块 51

5.3.2 视频播放模块 51

5.3.3 弹幕情感分析模块 52

5.3.4 视频评论模块 52

5.3.5 个人管理模块 53

5.4基于情感分析的视频弹幕系统的数据库设计 53

5.5 本章小结 57

6基于情感分析的视频弹幕系统的实现与测试 59

6.1 系统的开发环境 59

6.1.1 系统的软件配置 59

6.1.2 硬件平台参数 59

6.2 系统主要功能的实现 59

6.2.1 视频检索模块 59

6.2.2 视频播放模块 61

6.2.3 弹幕情感分析模块 63

6.2.4 视频评论模块 65

6.2.5 个人管理模块 67

6.3 系统测试 69

6.3.1 系统测试环境 69

6.3.2 系统功能测试 69

6.3.3 系统性能测试 72

6.4 本章小结 73

7 结论与展望 75

7.1 总结 75

7.2 展望 75

致谢 78

参考文献 80

攻读硕士期间发表的学术论文 82

CONTENTS

1 Preface X

1.1 Significance and Background X

1.2 Glance of Current Research Status X

1.3 Main Research Contents of Thesis X

1.4 Organization of Thesis X

2 Related Theory and Technology of Sentiment Analysis X

2.1 Common Technology of Text Sentiment Analysis X

2.1.1 Web Crawler X

2.1.2 Common Word Segmentation Technology X

2.1.3 POS Tagging X

2.2 Text Sentiment Analysis Method Based on Machine Learning X

2.2.1 Text Representation Model X

2.2.2 Text Feature Extraction X

2.2.3 Weight Calculating X

2.2.4 Classification Model X

3 An Sentiment Analysis Method of Barrage Text Based on SVM X

3.1 Problem Description X

3.1.1 Web Crawler X

3.1.2 Common Word Segmentation Technology X

3.2 Arithmetic statement X

3.2.1 Construction of An Emotional Dictionary Library of Video Barrage X

3.2.2 Barrage Text Preprocessing X

3.2.3 Emotion Feature Extraction X

3.2.4 An Sentiment Analysis Algorithm for Barrage Text Based via High Dimensional Hybrid Features Based on SVM X

3.2.5 An Sentiment Classifcation Model Combining the Objective Barrage Text X

3.3 Experiment Results and Analysis X

3.3.1 Experiment Performance Indicators X

3.3.2 Experiment Design and Analysis of Results X

3.4 Chapter Summery X

4 Requirement Analysis of Video Barrage System Based on Sentiment Analysis X

4.1 Requirements Description of the System X

4.2 Analysis Model of the System X

4.2.1 Function Model of the System X

4.2.2 Static Model of the System X

4.2.3 Behavior Model of the System X

4.3 Nonfunctional Demand of the System X

4.4 Chapter Summery X

5 Design of Video Barrage System Based on Sentiment Analysis X

5.1 Overall Structure Design of the System X5

5.2 Summery Design of the System X

5.2.1 Function Model Design X

5.2.2 Class Design X

5.3 Detailed Design of the System X

5.3.1 Video Retrieval X

5.3.2 Video Playback X

5.2.3 Barrage Sentiment Analysis and Visualization X

5.2.4 Video Forum X

5.2.5 Personal Management X

5.4 Database Design of the System X

5.4 Chapter Summery X

6 Implementation and Test of Video Barrage System Based on Sentiment Analysis X

6.1 System Development Environment X

6.2 Implementation of Main Module X

6.2.1 Implementation of Video Retrieval Module X

6.2.2 Implementation of Video Playback Module X

6.2.3 Implementation of Barrage Sentiment Analysis and Visualization Module X

6.2.4 Implementation of Video ForumModule X

6.2.5 Implementation of Personal Management Module X

6.3 Test of the Module X

6.3.1 System Test Environment X

6.3.2 System Testing and Results X

6.3.3 System Performance Test X

6.4 Chapter Summery X

7 Conclusions and Suggestions X

6.1 Conclusions X

6.2 Suggestions X

Acknowledgements X

References X

Achievements X

# 1 绪论

## 1.1 研究背景与意义

随着互联网的迅速普及，多媒体处理技术，网络传输、视频数据处理等相关技术快速发展，来自不同领域的视频数据正在以惊人的速度增长。在我国，视频网站、视频短视频网站、视频直播平台等视频播放平台因其内容丰富、内容新颖等特性在网民中迅速传播，尤其在移动互联网的时代浪潮下，大量的信息通过这类平台得到传播。在大数据时代，通过数据挖掘方法对此类平台交互数据传导的情感分析可从侧面反映出情感倾向[1]。

视频网站的产生已经有一定的历史，但是随着web2.0时代的推进，视频网站的展现形式开始变得丰富和新颖。视频网站不再是以往只有播放电视节目、电影资源的系统，而诞生了可以用户及时上传短视频的短视频feed流平台，还有可以进行用户视频直播的在线直播平台等等。而视频的内容也从简单的电视、电视节目，拓展到了设计动漫，直播，短视频，军事，电影，广告，综艺节目等等几乎无所不有的节目内容。视频平台已经成为人们生活中高度参与的一种资讯获取手段和休闲娱乐手段，这也进一步推动在线视频产业越来越火热，产品和功能变得更加丰富，研究学者对这一媒介的关注度也在逐渐升温。

随着视频平台的发展，视频平台给用户提供的交互功能也越来越丰富。从最开始用户只能观看视频，到后来用户可以对视频进行评论，打分；近年来，用户甚至可以使用一种交互感更强烈的手段对视频进行评价——弹幕。这些各类型的UGC数据和视频本身紧紧的绑定在一起，吸引了非常多的学者对这些数据进行研究和分析。这些研究包括了对视频评论的主题挖掘，情感分析，对用户观看记录和交互日志的行为分析等等。通过这写方面的研究，有利于挖掘视频本身的内容信息，情感走向信息，以及用户和视频之间的关联关系模型等。

目前，对web文本的情感倾向研究主要集中在电影评论、博客文章以及微博SNS等领域，而专门针对弹幕文本的情感分析还处于较为起步的阶段，相关的研究工作并不多。由于视频关联的弹幕信息越来越丰富，而弹幕这种和视频内容有很强的时间相关性，包含着丰富大量的信息，可以从侧面反映视频内容本身。因此，如何挖掘、分析并利用网民的弹幕语料，如何研究弹幕语料和视频内容本身的关系，并在此基础上寻找合理的分析模型，与传统的指数方法结合共同完善对视频的分析，给予用户更好的观看视频体验成为当下的热点问题。

## 1.2 国内外研究现状

1）文本情感分析研究

根据情感分类的方法，情感分析在微博、网络商品评论领域的应用研究可分为基于情感词的方法和基于机器学习的方法两大类，具体详情如下。

基于情感词的方法通常是现构建一个情感词典，词典中包含情感词所属的类别及对应的情感值，如正整数值表示正向情感，负整数值表示负向情感。Hu等对每一条用户使用反馈中各类别情感词的数量进行统计，在此基础上判断用户对产品的情感倾向。为了进一步量化用户评论的情感，Yu等把一条评论中所有的情感词的情感之平均数作为该条评论的情感值。Kim等考虑到否定词和双重否定对情感之的影响，采用成绩发判断某一条评论的极性。Yang等通过定义情感词权重词典、否定词词典、程度词词典及感叹词词典，综合考虑计算每条微博的情感指数。

机器学习方法主要是用用机器学习模型，对训练集的情感特征进行学习，提取系统输入输出之间依赖关系，从而应用于对测试集的分类判断。Pang等利用支持向量机，朴素贝叶斯、最大熵三种方法对电影评论进行分类，发现支持向量机的分类效果最好，而最大熵和朴素贝叶斯的分类效果相当。刘志明等研究发现，采用支持向量机的机器学习算法、信息增益的特征选取算法和TF-IDF的特征项权重计算方法，三者的结合对微博段文本的情感分类效果最好。Yu等则认为通过机器学习方法对新闻进行自动情感分类时，同时选取形容词、副词和名次作为情感特征项比选取单一词性的效果要好。

而从研究的文本表现形式上来看文本情感分析，文本在表现形式上有三种形式,即词语、句子和篇章。因此文本情感分析的研究也可以按照词语、句子和篇章这三个粒度级别来划分，其中本文主要针对弹幕这种句子级别的短文本，一下针对句子级情感分析方法来详细介绍。

句子情感倾向分析是对特定上下文中的句子中的例如句子的情感极性及其主体、客体、强度等主观性信息进行分析和抽取的任务,例如句子“大家觉得福克斯轿车品质还行但是外形绝对时尚”,该句中评论主体是“大家”,评价客体是“福克斯轿车”,“品质”,“外形”,“还行”和“时尚”表明这句话是正向情感倾向的,并且“时尚”的强度要大于“还行”。句子级情感分析可以算是篇章级情感分析的基础,在获得了各个句子的情感倾向信息之后,分析整篇文章情感极性的任务也就比较容易解决了。

姚天防等人在抽取主观句的主题及其属性时利用了领域本体的相关信息,对句子使用句法分析工具,从而得到句子中词语之间的句法依存关系来推测候选词情感极性。Gamon等人[2"]使用机器学习方向的基于朴素贝叶斯的句子褒贩分析方法完成了句子的情感倾向分析任务。蔡健平等人[21]以知网的词语语义相似度为基础提出了一个基于语义和语法距离度量的句子情感极性计算方法。王素格等人[22]以线性加权组合的方法为基础构造了一个用于判断观点句情感倾向的分类器,并且使用自己构建的情感词典中的情感词来表示观点句。熊德兰等人[23]提出了一种基于知网的句子情感倾向性计算方法,该方法融合了语义和语法信息,并且利用夹角余弦算法改进了语义倾向的计算方法。党蕾等人[241分析了修饰词极性和否定共享模式,将否定模式匹配与依存句法分析结合起来对情感修饰词的影响进行了研究,实验结果比较理想。李实等人[25]借鉴英文中的相关研究成果,分析中文语言的特点,在针对中文文本提出的规则和定义的基础上实现了对中文文本的特征挖掘。

2）文本情感分析在弹幕领域的应用

目前，基于视频弹幕数据的弹幕分析研究工作相对较少。有三个基于视频弹幕数据的视频分析技术可以对该问题起到一定借鉴作用。

其中，【？？】的文章是基于弹幕数据的视频片段自动标签技术，提供了一种提取视频片段中弹幕的关键字给这段片段打标签的技术，它的主要工作是从弹幕密集区域提取出核心的语料信息，不足是弹幕这种亚文化造成了很多密集弹幕高频词是客观词，无法直接分析出含义或者情感倾向。 基于弹幕的视频精彩镜头的检测方法【？？】主要根据弹幕内容以及数量的变化曲线分析视频中的精彩瞬间；它的主要工作是通过弹幕密度分析以及弹幕情感分析进行视频片段内容精彩程度的评估，它的不足是没有结合视频本身去分析，这样的推荐方法研究的重点有点跑偏，效果会有一点让人质疑。【？？】是针对弹幕文本进行了多位的情感倾向分析，利用多种可视化手段对弹幕情感分析的结果进行了可视化展示；它的主要工作是弹幕文本在细粒度上的情感分析和可视化，它的不足是研究结果并没有找到对应的应用场景。

## 1.3 论文的主要研究内容

本文以热门弹幕视频网站的弹幕语料分析作为切入点，根据弹幕密度分布对视频弹幕分片，然后依次对视频关联的弹幕文本进行情感计算，得出特点视频在时间轴上的情感波动曲线。本文对视频弹幕数据进行数据挖掘，分析弹幕所分析得出的情感倾向来反映对应视频片段的内容情节情感波动。并将这一情感波动信息提供给观看视频的用户，帮助用户了解视频剧情波动情况，更好的进行观看视频的决策。通过以上背景问题以及现状的分析，本文旨在深入视频弹幕数据情感分析的实际应用，力求寻找一个较好的结合方法，来实现一个可靠、有效的视频弹幕服务系统。为视频观看用户提供更好的观看决策信息，为相关领域的研究与发展提供具有重要创新意义的理论、方法与技术工具。

本文首先针对现有的web文本情感分析方法在弹幕文本领域的不足进行了全面的分析，随后提出了一种基于情感词典和SVM相结合的弹幕文本情感分析算法。实验验证算法的有效性之后，通过这一算法，得到了弹幕用户对视频内容的情感观点。随后，本文还提出了一种结合客观语料进行情感分析的文本情感分析方法，利用弹幕和视频之间的时间相关特性，计算出弹幕中的主观弹幕的情感值，映射给相关的客观词。通过海量的样本数据反复调整客观词的情感向量。在后续的情感分析工作中，进行过情感向量标注的客观词被加入到情感分析模型中，完善情感分析模型的完备性，提高情感分析的准确度。

本文在上述理论分析与实证研究的基础上，设计并实现了一个基于视频弹幕情感分析的视频弹幕平台。盖设计将严格遵守软件设计规范，进行系统的需求分析，分析系统用户需求，建立对应的静态模型和动态模型。然后再需求分析之后完成系统总体设计，即总体结构设计、类的设计、各模块功能设计、数据库设计和系统界面设计等。最后在需求分析和系统和设计之后，进行了系统的实现和测试。

## 1.4 论文的组织结构

本文一共分七章，各章节内容具体安排如下：

第一章绪论 论述本论文的选题背景以及意义，并调研分析了国内外对文本情感分析和情感分析运用于弹幕文本的研究现状，简单概括了论文的研究内容，并列出了论文的组织结构。

第二章相关理论以及技术介绍了本论文涉及到的相关概念和技术，包含本文提出算法的理论支持，情感分析技术的介绍，并对本文使用的SVM支持向量机做了介绍。

第三章股评倾向性分析研究，本章介绍了基于词典的网络新型文本的情感分析。首先,对于被清洗掉的短小的评论文本采用基于词典的分析方法。本文构建由基础词典、扩展词典的情感词典。通过对比实验可得情感词典与LDA主题模型进行结合后得到的情感特征的准确度更为精确。将得到的情感词典与LDA主题模型进行组合选取文本特征, 并按照其相应的情感倾向性计算规则，得到情感权值，情感均值，情感标准差这三种情感特征项。

第四章基于倾向性分析的股票资讯服务系统的需求分析。本章介绍了机器学习的情感分类基本流程。对待处理弹幕文本进行预处理后再特征选择部分，本文在支持向量机算法的支持下，分别针对不同的特征组合方式进行了分类效果研究，得到分类效果最佳的特征选择组合。第二部分算法研究针对于结合了客观弹幕语料的情感分析模型的研究。通过结合弹幕在时间轴序列上的密度排布，找出弹幕密集区域；通过主客观弹幕分类，主观语料输入分类器进行分类预测，客观弹幕进行高频短语提取，最终进行映射，构建一个客观情感弹幕词典。最终结合新生成的客观弹幕情感词典进行新的情感分类模型训练。在通过统一测试数据实验验证，结合客观弹幕的情感分类模型比普通的弹幕情感分类模型的预测准确度有明显的提高。

第五章针对视频弹幕情感分析系统的设计。本章将本章将从系统构架设计、系统概要设计、系统详细设计和数据库设计四个方面对系统设计进行讨论，为系统的实现打下基础。

第六章基于倾向性分析的股票资讯服务系统的实现与测试本章将基于股票评价的股票资讯服务系统系统的实现环、实现的核心代码、功能截图以及测试过程进行详细的展示，以描述系统的实现与测试。

第七章总结与展望 总结全文，并以本文对弹幕文本的应用研究为出发点，指出了网络短文本在情感分析领域进一步研究和发展的方向。

# 2 相关理论及技术

本章主要介绍了了本情感分析的基本流程，以及所设计的相关理论知识体系与相关技术。如文本与处理、文本表示、特征提取与选择、特征权重计算、情感词典、分类算法、分类起性能评估等理论基础。

## 2.1 文本情感分析常用技术

根据所使用技术的不同，情感分析的技术通常可以分为两类：一类是基于情感词典的额方法，其主要步骤是对文本进行文本与处理后，与以构建的情感词典进行对比，获取到相应的情感词，继而计算该文本的额情感倾向。另一种是基于机器学习方法，即首先对语料库进行训练数据集和测试数据集的划分。经过文本预处理后，得到文本特征的矩阵化表示，用文本分类算法完成对文本情感的倾向判别。基于机器学习的情感分类的基本过程如图所示。基于机器学习的情感分析可以分为两部分：训练分类起和训练模型的测试。在训练分类起部分，首先进行文本预处理和特征选择以及训练集文本的表示，并且通过分类算法训练特征向量后，获得分类模型。在测试训练模型部分，通过测试集文本，特征选择，特征表示，然后在训练模型的分类上利用测试集对训练模型进行测试。因此，文本预处理，文本特征选择和表示，机器学习分类算法是文本情感分类研究中必不可少的部分，下面将对以上部分的理论和技术方法进行更详细的介绍。

2.1.1 网络爬虫

网络爬虫的工作原理是从一个或若干初始网页的链接开始进而得到一个链接队列。伴随着网页的抓取又不断从抓取到的网页里抽取新的链接放入到链接队列中，直到爬虫程序停止。其工作流程如下[26]： 1）首先选择需要抓取的网页，作为种子URL； 2）将需要抓取页面的URL组成队列，即待抓取URL队列； 3）从上一步由URL组成的队列中依次取出待抓取的URL，解析其中的DNS特征，将它们对应的网页下载下来，并按照事先设定好的结构存储进数据库中。已经抓取过的网页URL会被放入已抓取URL队列。 4）对已经分析过的URL队列，分析页面中包含的其它URL信息，如果出现没有抓取过的页面，则需要将新的URL放入待抓取URL队列中进入下一个循环。 一个通用的网络爬虫用的框架如图2-1所示：

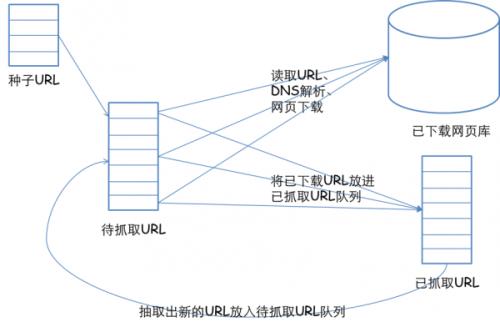


图2-1 网络爬虫架构图

2.1.2常见分词技术

分词是对文本处理和文本分类任务的第一步，中文分词与英文分词在分词过程中会有很大的不同，英文文本中英文单词间有空格对单词进行切分，所以在英文文本中空格起着切分英文词汇的作用。但是在中文分词中这个简单易行的优势将不会得到体现。因此在没有对中文文本进行相关处理时就应先对文本进行分词处理。

目前主要有三类基于分词的研究方法。

其中基于词典的字符串匹配的分词方法根据词典、顺序扫描以及匹配等展开研究，其优势在于实现相对容易，切通过匹配的方法能准确辨认出文本出现的词语，但是未登陆词的处理工作没有进行研究，并且词典之间犹豫标准的不同意会出现词汇歧义的问题。由于基于顺序扫描与匹配的过程，其匹配计算量较大。基于统计的分词算法主要根据相邻的字的共现频率，通过其共现次数确定其关联关系。崇勇的统计模型有互信息、N元文法模型等。基于统计的分词方法的优势主要体现在消除歧义和未登录词的识别两个方面。

在对中文分词的研究过程中，基于统计的中文分词其主要思路时词汇在上下文中，统计字与字之间的词语组合，两个字在同一文本中的共现频率高，其构成一个词的可能性就越大。基于统计的中文分词其优势主要体现在不受限于词典与文本的相关规则，但由于要用语料库对其进行模型构建与测试，其构建过程相对复杂，切分词的京都会收到训练文本等相关因素的影响。进行基于统计的中文分词方法常用的统计模型有：互信息、N元文法、HMM(隐马尔可夫链)、最大熵等。

本文将选用NodeJs中文分词中的Jieba分词开源库对文本进行中文分词。Jieba分词的基本原理主要是基于Trie树结构对文本进行扫描，并输出分局中词组成像一个的有向无环图结构，根据动态规划算法查找最大概率的路径。Jieba分词采用基于HMM模型使用Viterbi算法解决未登录词的识别问题。

2.1.3 词性标注

词性标注简称标注，是指对文本进行分词处理之后，对其相应的词汇进行词性的标注的过程，即标注分词后的词汇是动词、名次、形容词还是其他词性，然后将其词性（Part-of-speech，POS）信息分配给相应单词，并用标签进行标识。目前比较常用的词性标注算法如图。

基于规则的方法主要原理是在规范的规则下，对语料进行词性标注。其优势是标注过程相对简单且易懂，当爱方法需要人工执行消除器一规则，随着语言形式与新兴词的出现，人工制定的歧义消除规则不但任务繁重且难以保证规则之间是否存在歧义。所以方法在处理新兴词汇和网络词汇是较为困难。

基于统计的方法其主要原理是通过分类算法的方式来自行推衍文本中的信息与规律。其优势在于该方法主要根据对数据的统计，通过算法自行归纳规则，无需人工制定标注规则。但是由于是基于数据的统计，其总结推算出来的是一种基本似然规律，无法对其进行定性的总结归纳。由于语料库的不全面性会导致词性判断的片面性，因此要考虑由于稀疏数据造成过度拟合等问题。

根据规则与统计的方法的利弊，结合规则与语料库统计的方法形成基于规则与统计相结合的方法。其主要原理是对词性进行标注时，使用规则的方法对简单语句中的词汇进行词性标注，节省资源。若使用规则的方法处理较为困难时，再使用基于统计的方法来处理相应词性标注的问题。

在中文情感分析中，中文语句会出现很多不含邮情感色彩的虚词，为了使情感词典的构成尽量简化与计算量的减少，我们着重对语言特征与属性特征的词汇进行词性标注和收录。在词性标注中，标注的词性只具有参考作用，其并不具有绝对的准确性。对于上下文中标注的词性存在歧义问题，也有相关的学习方法对其研究提供了较为夯实的基础与清晰的思路，例如隐形马尔科夫模型、基于转换的学习方法。虽其取得了相对较好的研究效果，仍然还需要广大学者对其继续研究。

## 2.2 基于机器学习的情感分析方法

2.2.1文本表示模型

文本主要是由字、词、标点符号以及相应的句式构成。为了确保文本分类的高效性与准确性，文本形式话表示就显得格外重要，因此要找到技能标识不同种类的文本并能准确的反映文本内容的方法是进行文本分析的必要步骤。

文本的形式化表现就是从文本中抽取能表示文本特征的信息，并通过规则对其加以形式化表现。目前通常的文本表示方法有向量空间模型（VSM）、语言模型、后缀树表示法以及本体表示法等。本文将采用向量空间模型作为文本的形式化表现方法。向量空间模型（Vector Space Model，VSM）的主要原理是将文本应设成一个向量的点，文本的每个特征由向量的每一维来表示。通过对文本进行分词和特征提取处理再进行计算。

空间向量模型的主要原理如下所述：

1）文档（Document）：文档是VSM中处理的对象单元。

2）特征项（features term）：在VSM中，最小不可分的单元就是特征项，特征项包括字、词、词组和短语等。

3）特征项的权重（feature weight）：每个文档中的特征项ti标识在文档中重要程度的值，即权重wi。通过VSM，就可以将一个吻昂表示为n维空间的一个空间向量。

4）空间向量模型：给定一个文档D(t1,w1;t2,w2;…;tn,wn)，D满足以下两条约定：其一，不考虑文档的内部结构；其二，文档中各个特征项没有重复。

2.2.2 文本特征提取

对文本进行建模之后，提取相应的特征项就可以构成初始特征集合。如果每个特征词代表一个特征维度，那么目标文本的特征维度会非常高，因此，需要对文本的特征进行降维。此外，由于自然语言文本的特征本身比较稀疏，如果把所有的特征都考虑在内，进行特征向量的分类，那么计算会非常复杂，效率低且效果差。由上可见，如果想提升文本情感分析的正确率，提高算法的计算效率，就必须对文本的特征进行降维，并去掉出现频率很低的特征。在对分类器进行训练时，我们应该保留关联度高、出现频率高的特征，而对分类效果影响不大的特征进行过滤。通常情况下，我们使用常用的特征提取和特征选择方法来对建模后的文本进行降维，并以此简化算法的复杂度，提升分类器的分类效果[35]。

文本特征提取的方法很多，针对不同的应用领域往往有各自的优化方法。但是它们的基本思路往往较为统一，就是构建一个从复杂的维度向简单的维度进行转换的方法，并通过特征向量维度的降低实现算法复杂度的降低，提升系统的计算速度与分类器分类的准确性。目前，最常见的特征提取方法包括主成分分析、非负矩阵分解以及潜在语义索引三种。

特征的选择方法，是指按照一定的标准，从原始的特征集合中挑选出最具有区分度，对分类效果影响最大的特征项。这需要我们能够减少对文本分类没有帮助的特征，减少无用特征对分类算法的干扰，从而进一步降低向量空间的维度。特征选已经被广泛的应用在各个领域，降低了文本分析的计算复杂度，提升了特征对文本语义描述的相关性。目前，常见的特征选择方法包括互信息(Mutual Information)、𝜒2统计量(CHI)、信息增益(Information Gain)、以及倒排文档频率(Inverse Document Frequency)和文档频率(Term Frequency)等等。

2.2.3 权重计算

在向量空间模型表示文本中，我们首先会根据各个特征项对表示文本的贡献度以及对文本分类能力进行特征选择，而我们会使用特征权重来量化这种分类能力。特征甲醛就是赋予各个被选择出的特横的权值的过程，其常用的方法有以下三种：

1）布尔函数

布尔函数的数学表达式如公式：

（2-1）

为文本中的第i个特征项的权值，表示文本中的第i个特征项在该文本中出现的次数，即词频。从公式可以看出，该方法不考虑特征项出现的次数，仅仅简单笼统的根据出现与否特征性赋予1或0的权重，因此其忽视了词频对文本分类的影响，未能很好的体现各个特征项的区分性。

2）开根号函数

开根号函数方法就是将特征项的词频的算术平方根作为该特征词的权重，其公式如：

（2-2）

该公式各项的表示同前面的公式，虽然该方法较布尔函数的方法有了改进，不是笼统的0，1权重，但是该方法得到的权值是随着特征项的词频增加呈单调增加趋势的，这违背了特征选择的原则。开根号函数方法在表示特征对文本的贡献度的效果方面也不是很好。

3）TF-IDF方法

TF-IDF方法实现技术简单且高效，因而其得到了广泛的应用，计算公式如下所示：

（2-3）

TF称为词频，是特征项t在文本中出现的频率，IDF称为逆向文档频率，表示含有特征项t的文本在整个文本集中的分布情况。这种方法既考虑了特征项的词频，又考虑了该特征项的反文档频率，TF表示的是该特征项在单个文本中出现的频率，IDF则表示的是该特征项在整个文档集中的出现频率。TF说明了特征项对该文本内容的表示能力，IDF则说明特征项对文本分类的贡献能力。

本文根据特征不同选取不同的特征权值，如否定词特征，程度词特征选取布尔函数赋值，情感词特征、词聚类特征等选取TF-IDF函数赋值。

2.2.4 分类模型

用于分类的算法有很多,如:逻辑回归分类法、贝叶斯分类法等。它们的共同特征是分类误差最小化。 但该特征存在两个问题: 一是不能判别划分数据的可信程度; 二是可能得到的解有多个甚至无穷个, 不能判定所有解的优.劣 。 为解决上述问题就出现了支持向量机(SupportVlectorMachine, SVM) 。 SVM的核心思想是:使各点的间隔最大化。

通过其数学公式推演将其转变为如下形式:

（2-4）

SVM方法是通过一个非线性映射p，把样本空间映射到一个高维乃至无穷维的特征空间中（Hilbert空间），使得在原来的样本空间中非线性可分的问题转化为在特征空间中的线性可分的问题．简单地说，就是升维和线性化．升维，就是把样本向高维空间做映射，一般情况下这会增加计算的复杂性，甚至会引起“维数灾难”，因而人们很少问津．但是作为分类、回归等问题来说，很可能在低维样本空间无法线性处理的样本集，在高维特征空间中却可以通过一个线性超平面实现线性划分（或回归）．一般的升维都会带来计算的复杂化，SVM方法巧妙地解决了这个难题：应用核函数的展开定理，就不需要知道非线性映射的显式表达式；由于是在高维特征空间中建立线性学习机，所以与线性模型相比，不但几乎不增加计算的复杂性，而且在某种程度上避免了“维数灾难”。

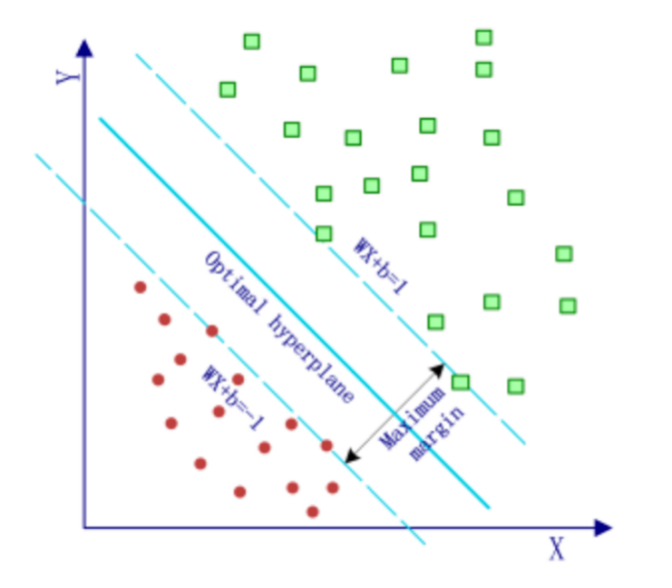


图2-4 支持向量机原理图

## 2.3 本章小结

本章节首先介绍了文本情感分析的一般流程：文本预处理、特征表示、机器学习、文本分类等步骤。接着详细介绍了文本预处理的步骤中，中文分词和词性标注等。文本预处理后的文本表示模型将采用向量空间模型。2.2介绍了几种传统的特征提取方式，如文档频率、信息增益法等并分析了方法的利弊。2.3介绍了分类算法SVM的基本原理。

# 3 基于SVM的弹幕文本情感分析算法研究

上一章主要介绍了本文涉及到的基本概念与关键技术。从本章开始，将介绍一种针对弹幕文本的情感分析方法。本章首先对研究的主要问题进行描述，指出现有web短文本情感分析方法的问题。然后描述了弹幕文本的挖掘过程，并就这些问题提出了基于SVM的高维混合特征弹幕文本情感分析算法和结合客观弹幕的情感分析模型，分别进行针对性的算法研究。最后将本章提出的算法与常用SVM分析算法进行实验对比。

## 3.1 问题描述

3.1.1弹幕文本的特点

目前主流的文本情感分析算法在对弹幕短文本进行分析时效果往往不够理想。这主要是由于弹幕这种形式的短文本和普通的网络评论文本有着显著的特征差异。这种差异主要体现在以下几个方面。

1)内容形式较短。

在弹幕这种基于web视频流媒体之上存在的新型评论文本形式中，单个文本的长度相较于传统的网络评论文本有显著的差异。弹幕的长度整体偏短，绝大部分弹幕长度在一句话以内。弹幕长度短会导致可提取有效信息减少，影响情感分析结果不够准确。

2)表达形式多样。

核心的特点就是表达形式多样，相对传统评论更加不受约束，这一点深受弹幕的主要受众们（主要涵盖了青少年和年轻人群）的喜爱。为了表述强烈的情感寓意，弹幕中会高频的出现字符表情，这种形式的目的是在表达中凸显俏皮和幽默的气氛，用来达到加强情绪表达和观点阐述的效果。还有很多弹幕为了简洁明快的表述观点，会用一些缩略用语来代替标准但冗长的原语句。这种表达方式在视频情节的一些高潮区间会大量出现，目的是表达出观众们对特定剧情的强烈态度和情绪。

还有一种很重要的弹幕表达形式是“隐喻”，这种表达方式的文本往往是不含主管情感倾向的客观语料，但是实际上文本中包含了强烈的情感倾向信息。本文将常见的包含情感倾向的客观弹幕语料形式总结如下。

a.象形文字/符号。

例子1: “orz” ，形状像一个面朝左跪在地上的小人，表示强烈的悲伤的氛围。

例子2: “23333”，意思是“啊哈哈哈哈”，模拟大笑的声音，表示开心的心情。

b.借用一些典故的用语反讽，语义上的情感倾向和实际倾向恰恰相反。

例子1: “xx感人”，语义上是说xx让人感动，而实际的用意则是负面态度，暗指xx这件事情让人不够满意。

例子2: “我从未见过如此厚颜无耻之人”，原文是来自三国演义中王朗的台词，原为贬义；后在弹幕中则是褒义，形容某人做某件事情的行为方式很巧妙，这是一种亲昵的表达方式。

c.还有一些出现在弹幕中的短语模式，表达出的情感倾向需要根据具体上下文才能确定。

例子1: “膝盖”，当弹幕中出现“跪了，把我的膝盖送给你”时表示敬佩的意思，表示褒义；当弹幕出现“膝盖中箭”时，表示的是很受伤的负面情绪。这种一词双意的状况也是弹幕文本情感一种很特别的表达形式。

d.一部分弹幕是一些常用弹幕文本的缩略语。

例子1:“颜表立”，全文是“用颜色表达立场”，是动漫视频中常见的剧情内容；

例子2:“火钳刘明”，全文是“这个视频火之前，我在这里留一个名字”；这里不仅缩略用语，而且还有使用同音词代替。

如上可以发现，各种各样的客观语料组成了弹幕这种新兴的网络文本载体。这些文本包含了及其大量且重要的情感倾向信息，我们在对弹幕的情感分析过程之中非但不能将其视为噪声忽略，还应该着重加以利用，来改进我们的弹幕情感分析模型的预测准确率。

3)术语的特定性。

弹幕在产生初期使用者的圈子局限于动漫爱好者，由此造成了会在弹幕中出现很多动漫圈特有的术语，如“卡哇伊”，“宅”，“前方高能”。之后在弹幕发展的过程里，也不断的产生着很多特有的术语。如“蓝瘦，香菇”，“膝盖中箭”等等。在分析过程中对专业术语的缺乏显然会导致情感分析结果不够准确。

3.1.2 方法概述

基于文本的情感分析涉及到自然语言处理、文本挖掘、数据挖掘、机器学习等多个领域。如图3-1所示，一个基本和典型的情感分析问题的解决，包括下面几个步骤:

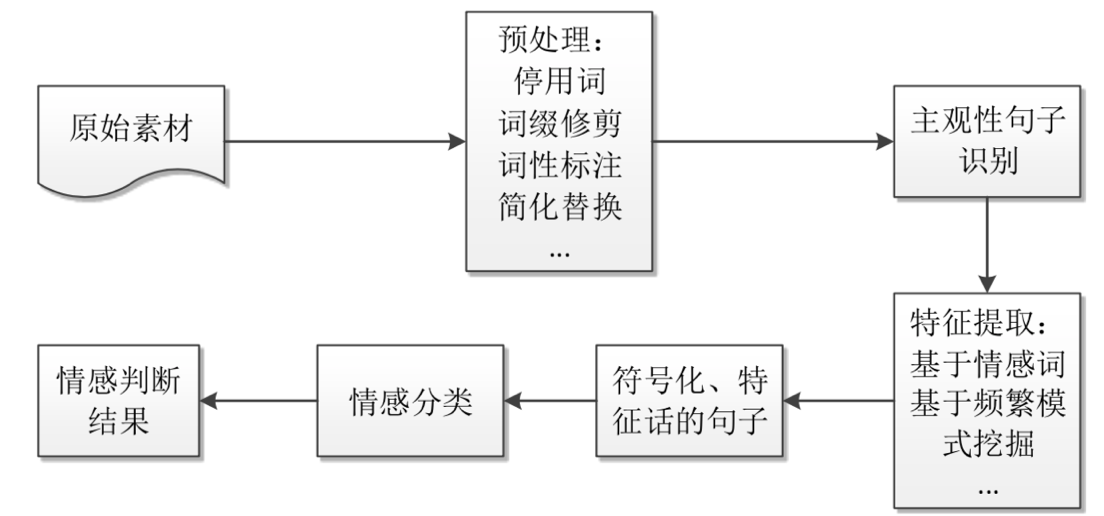


图3-1 文本情感分析过程图

弹幕文本是一种典型的web短文本，针对它的词汇和结构特点，相关研究做出了以下几种改进方式。

基于词典扩充的web短文本倾向性分析方法首先针对web评论中的专业词汇进行情感词典扩充，然后结合对多种特征选择算法的实现与分析，提出了一种改进的卡方统计法，该算法不仅可以保留卡方统计方法对具有区分度特征识别的优点，而且能够避免低频特征项对系统整体的干扰。实验证明该方法在web短文本的分类中能够较大地提升查准率，但由于缺乏对篇章结构的分析，该方法在查全率上提升较低。

基于机器学习的中文微博情感分类实证研究的研究表明，采用一种结合支持向量机的机器学习算法，结合信息增益的特征选取算法和词频逆文档频率特征项权重计算的组合方法，对微博短文本的情感分类效果最好。但是弹幕文本作为一种新类型的网络评论方式，弹幕写作简单随意，并无严格的语法要求，这导致该方法的情感分析方法不能直接适用于弹幕情感研究工作之中，只有一些借鉴的作用。

基于以上的弹幕文本的特点，本文提出了一种基于SVM的高维混合特征的弹幕情感分析算法。并随后以此模型为基础，构造了一个结合客观弹幕语料的SVM情感分析模型。该方法首先使用半监督的方式构建弹幕情感词典库，然后对传统的词性特征提取方法进行了改进，对语料中的所有语句，按照本文所设计的规则进行分析，找出他们的情感极性，并分析所有带有情感极性的词汇的词性特征。随后，通过本文设计的词性匹配算法，根据倾向性分析的准确率，与词汇在文本中的出现比重提取词性的情感特征。最后将词典语义分析方法与SVM模型相结合，得到最终的倾向性识别结果。实验证明，本文提出的改进分析方法能够有效提高针对弹幕文本的倾向性分析查准率和查全率，效果明显。下面，将对本算法进行详细的描述。

## 3.2 算法描述

采用传统的文本特征作为支持训练机的特征输入，生成的支持向量模型可以用来对普通web文本进行倾向性分类。但是针对弹幕文本有诸如文本较短，表达方式多样等特有的问题，传统的特征工程在这一场景下性能会大大下降。针对如上两种问题，我们在此章提出了两种改进基于弹幕的情感分析的算法。1）一种基于SVM的高维混合特征的弹幕文本情感倾向性分析算法；2）结合客观弹幕文本的机器学习情感分类模型。

3.2.1 视频弹幕情感词典库的构建

本文知网的"情感分析用语词集"和台湾大学的NTUSD词典以及正面词汇、负面词汇和程度副词组合为通用情感词典。知网发布的"情感分析用语词集"中有中文情感分析用词语集,并且台湾大学发布的NTUSD词典中包含2810个正极性词语和8276个负极性词语。

首先筛选并去除知网的"情感分析用语词集"和台湾大学NTUSD词典中存在许多情感极性相反的,或者情感意义比较多的词汇,然后将知网(Hownet)里面的正面评价词语、正面情感词语和NTUSD词典的positive词典进行词语去重,然后对其进行组合,成为通用基础词典中的积极情感词典。把知网的负面评价词语、负面情感词语和NTUSD词典的negative词典进行词语去重,然后对其进行组合,成为通用基础词典中的消极情感词典。

通过使用这种扩充策略,对HowNet和NTUSD词典进行相应的关系映射,将所得词汇加入扩充词典中。通过使用构建策略,能有效提高工作效率,并且能増加许多有效的词汇对后续部分的词典扩展也有很好的借鉴意义。最后再把通用情感词典中重复的词语剔除,成为最终的通用基础情感词典。

3.2.2弹幕文本预处理

随后，需要完成弹幕文本预处理。预处理即过滤文本中不具备实际意义的词汇，得到任务模型所需的文本格式。其中涉及到的具体任务包含中文分词、词性标注、停用词去除等。具体算法流程图如下：

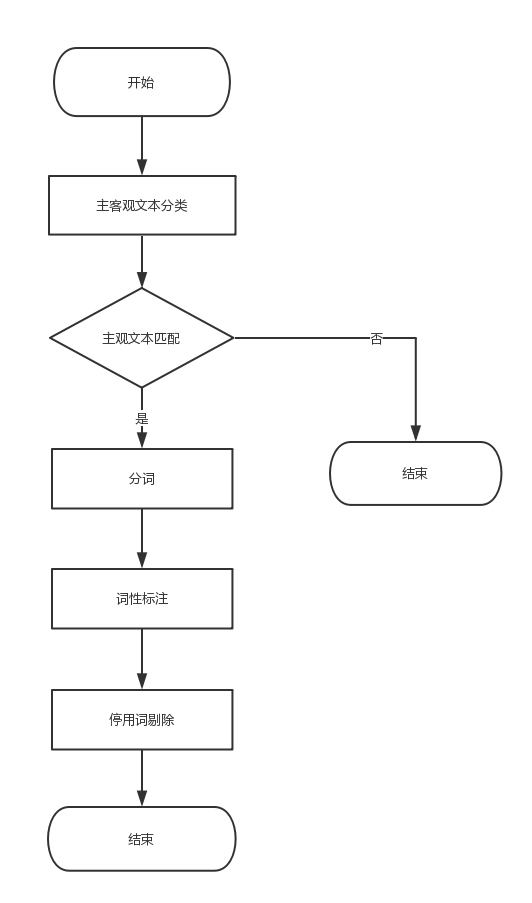


图3-2 弹幕文本预处理流程图

3.2.3 情感特征提取

接下来需要对预处理完成的弹幕文本进行情感特征提取。本算法选定了经典的文本情感特征项，诸如情感词特征、否定词特征，词性标注特征，ngram特征，还新加入了词聚类特征，希望针对弹幕这种短文本特征稀疏的问题加以改进。具体特征项描述如下：

1）情感词

在对文本情感分类时，往往文本中含有的少数带有情感倾向的词汇最直接表现文本情感的倾向性。如正向词汇“高兴”和负向词汇“难过”。由于中文词汇的复杂，情感词汇非常丰富，多位形容词、副词等。本文选择四个情感词典进行情感特征选择。其中包含整理好的Hownet、NTUSD、大连理工大学的本体词汇。其中由于前两者Hownet、NTUSD词典并没有标注情感词的情感极性。所以我们将正向词汇的得分设定为1.0，负向词汇的得分为-1.0。在情感词典特征上我们采用下面四个规则进行情感分数的计算：

规则1 分别计算情感文本中的正向词、负向词的数量；

规则2 分别计算情感文本中的正向词、负向词的得分总数；

规则3 分别计算情感文本中的得分最大正向词、负向词的分值。

规则4 分别计算情感文本中的最后一个情感词的分值。

2）否定词特征

含有主观倾向的语句往往有很明显的否定词。与传统文本情感分类不同，“不”“没”等否定词不再作为停顿词被删除。在句子里“不”或“没”的否定范围时“不”或“没”的全部词。一个词在不在否定范围内对正确情感分类产生很大影响。例如：他一直没上班/他没一直上班；你没天天学习/你天天没学习。本文采用否定特征是以句子出现否定词为否定特征的开始直至句子结束都加上否定标记。并且记录否定词的个数也作为否定特征的一部分。

3）词性标注特征

常见的分词系统的词性标注的粒度能达到：名次、动词、形容词、副词等。本文选用开源的Jieba分词系统。它能够将词性标注力度更为细分：例如名词可以分成人名、地名；形容词可以分为复形词、名形词、形容词性语素，形词词性惯用语。

4）n-gram特征

对于给定的文本都可以看作是长度不同序列的集合。在这些序列中相邻的N个字或词成为n-gram，n-gram算法的基本思想是通过一个大小为N的滑动窗口将文本内容进行切分，形成长度为N的片段序列，每个片段序列成为gram。使用n-gram特征，尽可能获取有限长度短文本的未登录情感词和情感信息。

传统分词技术对于短文本的分词存在的明显缺陷甚至可能会改变原有评价对象。本文鉴于此类情况增加n-gram特征：对于1-gram是单个的字或词对于特征的选择并没有多大的意义，所以本文选择从2-gram开始，但超过4-gram同样没什么意义。

5）词聚类特征

词向量具有良好的语义特性，是表示词语特征的常用方式。考虑到很多字、短语表现形式不同，但是表达的意义相同。采用词聚类的方法能够有效的避免处理相近的词语多余的计算量。Word2vec是Mikolov等所提出模型的一个实现，可以用来快速有效的训练词向量。可以吧对文本内容的处理简化为向量空间中的向量运算，计算出向量空间上的抗四度，来表示文本语义上的相似度。

word2vec包含了cbow和skip-gram两种训练模型。本文在skip-gram模型完成了词聚类构建了词聚类词典。在这个模型中，给出了语料库的词w及其上下文c，在给定条件概率和语料库Text，去求参数使得属于语料库的概率最大化：

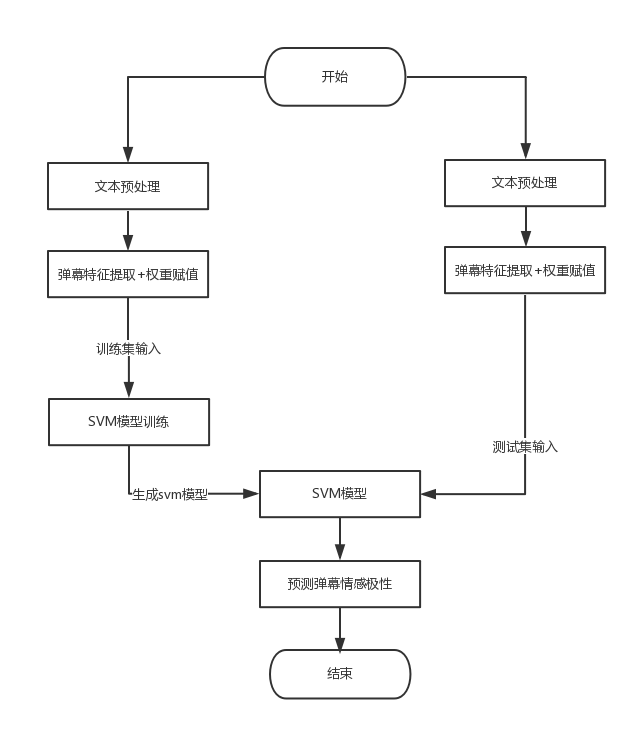
(3-1)

在这里C(w)代表一组单词w的上下文。本文使用该工具运用在收集到的语料库上聚100类后得到1533个基元。

提取完如上特征之后，还需要对特征依次赋权值。

3.2.4基于SVM的结合高维混合特征弹幕文本情感倾向性分析算法

基于SVM的高维混合特征的分类算法主要是先对待分析文本语料进行文本特征提取，通过这种方法生成一个混合特征的向量，然后将该向量作为SVM的输入并进行学习，在通过训练好的SVM分类器进行文本分类。该模型又文本预处理、词典语义分析、机器学习分析三个大的部分组成。第一部分主要是进文本预处理，使用Jieba汉语词法分析系统对所有文本进行中文分词、词性标注和去停用词操作。第二部分是提取弹幕文本中的多个特征向量，得到特征向量数组。第三部分是通过使用支持向量机的分类方法对文本情感极性进行分类。算法模型图如下图所示：



#### 图3-3 基于SVM的高维混合特征情感分析模型流程图

在流程的最后一步，进行SVM高维混合特征情感分类器的训练工作。通过将上一步中生成的弹幕文本特征向量作为svm的输入，训练并生成支持向量模型。接着使用第一步中生成的测试数据输入模型，进行验证测试。通过调整特征维数对情感分类问题的影响，以期得出最佳的特征维数。由于位数太低会造成区分能力差，不能够准确分类等问题；而维数太高又会出现数据系数、过拟合等问题，继而降低分类效率。因此，关于特征位数的选择对情感分析的分类问题也同样起着至关重要的作用。

与此同时，针对弹幕短文本提出的词聚类特征，是否选择此特征对模型预测准确度的影响需要通过实验得到验证。

利用训练集和测试集得到最佳的分类算法和特征维数后，还需要对语料库中的测试剂进行实证测试。将测试后的得到的分类预测结果和人工标注的标签进行对比，得到基于整个语料范围中，在最佳特征位数的条件下分类准确度最高的分类算法。也需要证明本文提出的词聚类特征针对弹幕短文本进行情感预测是否有改进效果。

3.2.5结合客观弹幕文本的情感分类模型

本章在前一小节基于SVM的弹幕文本情感分类模型的基础上，建立了一个高维混合特征的情感分析的模型，实现了对弹幕文本实现了情感分类的目的，达到了一定的准确率。

与此同时，我们发现弹幕中超过一半的语料都是客观语料，用户善于用“隐喻”的方式去基于客观语料去表达自己的情感和观点。基于这个状况，弹幕文本中的客观语料就不再适合被当作噪声所剔除，反而应该被当作一个重要的数据来加入到研究的流程当中来。

本文已经在3.1.1节中对弹幕这种特殊的短文本形式的特点做了详细的分析。针对这些客观弹幕语料，本文提出了一种结合客观语料的弹幕文本情感分类模型，具体流程如下。

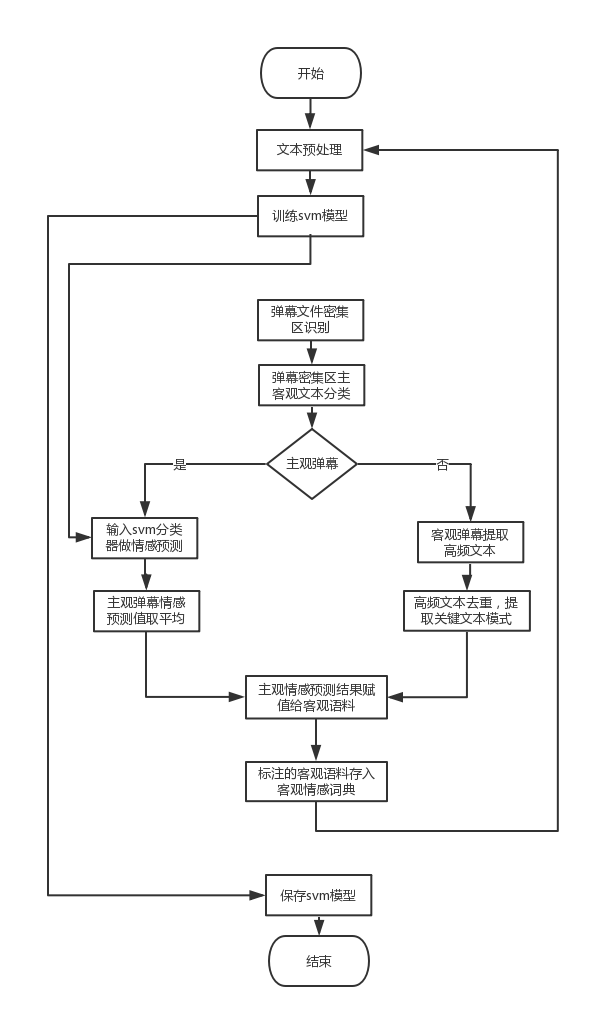


图3-4 结合客观弹幕文本的情感分类模型流程图

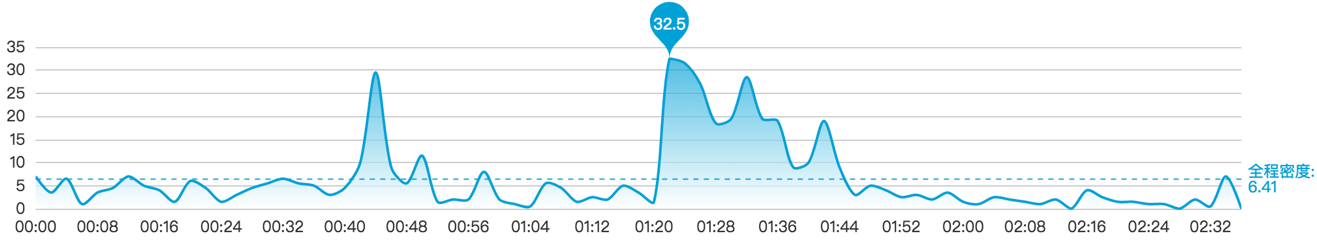
流程的具体细节描述如下，可以分为六个主要的步骤。

1）初始svm模型生成

这一模块是作为整个模型的前置模块，基本流程是基于上一节中我们建立的针对弹幕短文本的svm情感分析模型的训练过程去执行的。目的是根据弹幕主观语料先初始化一个基础版本的svm情感分析模型。

2）弹幕文件密集区域识别

这一部分是整个模型核心的一个模块。我们希望通过分析弹幕文件中，每一个时间点的弹幕密度，识别在当前弹幕文件中弹幕密集区域的区间集合。



#### 图3-5 弹幕密度在时间序列上的可视化折线图

上图是一段23分钟的视频弹幕在时间轴上的密度分布折线图；通过密集区域识别，获取到的密集区域数组。并分别对每一个密集区间的弹幕进行主客观分类。

3）对主观语料情感倾向分类

基于上一章训练好的初始svm弹幕情感分类器之上，我们将当前密集弹幕时间区间的主观弹幕部分输入分类器，并获取主观弹幕的平均情感倾向。输出的结果和对应密集弹幕时间区间相关联，并存入文件中。

4）对客观语料提取高频短语模式

当前密集弹幕时间区间中的客观语料部分，我们希望提取其中的高频出现的文本模式，并对其进行情感赋值。我们首先对客观语料按照出现频率由高到低进行排序；针对相似的弹幕语料进行关键文本模式的提取；去除重复的语料文本；最后提取排序靠前的短语作为结果输出，存入文件。

5）对前两步产生的情感分类结果和短语模式映射

按照对应关系，将同属一一个密集弹幕区间的主观语料情感值和客观语料高频短语模式进行映射，将映射结果存入新建的客观情感语料库。

6）结合客观情感语料库重新训练svm模型

如题所述，我们结合新生成的客观情感语料库重新训练针对弹幕文本的svm情感分类器。与初始化svm分类器时不同的是，我们在文本预处理阶段，不再将客观弹幕作为噪声数据剔除；而是将客观弹幕结合客观情感语料库进行情感分析，作为训练svm模型时一个重要的特征向量。重新训练svm模型，供随后对弹幕进行情感分类使用。

我们这样做的初衷是，弹幕文本密集的区域往往是由于对应时间点的视频存在一些剧情的高潮，引发人们通过发弹幕的方式去进行情感共鸣的回馈。所以我们有理由认为，在这些时间区间中的弹幕大概率是针对相同的目标事件才产生的。所以我们可以尝试对该区间内的含有情感的主观语料进行情感预测，并将情感预测值赋给遇见内高频出现的客观短语（长短不定的文本模式），并保存在我们新建立的客观情感语料库。

关于该模型的实验参比模型，为了验证结合客观语料的svm情感分类器模型的预测效果是否有一定的提升，我们选用上一章中基于六种情感特征的不包含客观语料的svm情感分类器模型作为参比模型。

## 3.3 实验结果与分析

3.3.1 实验性能指标

实验结果的评价指标应该考虑到多方面因素的影响，保证客观、公正的评价。一般的情感极性分类性能判定又准确率(Precision)、召回率(Recall)、F值三种评价指标。假定A代表分类算法能够将带测试文本准确的分道该极性类别的文本语料数量；B代表分类算法对文本错误的分道该极性类别的文本语料数量；C代表分类算法未能将文本准确的分道该极性类别的文本语料数量。

准确率，也叫查全率，记为P,顾名思义，该指标是指判定分类的准确性。准确率是判定该类别极性书与所有判定为该类别极性书的比值。其计算公式如下：

(3-2)

召回率，也叫查准率，记为R，该指标主要是为了判定分类模型的完备性，是正确判定为该类别极性书与本应该判定为该类别的比值。其计算公式如下：

(3-3)

F值记为F，由于实际测试中，当准确率高时召回率低，当召回率高时，准确率低。两者一定程度上呈相互制约的关系。为了寻求两者的平衡点引入了综合度量指标F值作为两者的调和平均数。其计算公式如下：

(3-4)

3.3.2 实验设计与结果分析

本章的实验主要针对两个方面，一个是基于SVM的高维混合特征弹幕情感分类实验，验证和经典的特征提取训练出的模型的情感分析准确度是否有提升；另一个则是结合客观语料的弹幕情感分类对比实验与传统的剔除客观语料的情感模型的比较。

1）基于SVM的高维混合特征弹幕情感分类实验

本章的实验主要针对的问题是，基于SVM的高维混合特征的弹幕文本情感倾向性分析算法与基于传统特征提取的SVM算法的优劣比较。

此次试验准备基于情感词特征、程度词特征、否定词特征，标点符号特征，n-gram特征这六种基础特征，以情感词特征作为基本特征，分别加入其他五种特征进行组合和试验。

为了研究基于SVM的高维混合特征的弹幕文本倾向性分析算法在弹幕领域内是否对最终的分析结果有所提升，本文选取了5500条弹幕文本作为样本进行测试。其中标注情感正向1540条，情感负向1760条，情感客观2200条。

其实验结果如下表所示：

表4-1 基于SVM的高维混合特征的弹幕文本情感倾向性分析表

|  |  |  |  |
| --- | --- | --- | --- |
| 特征组合 | 查准率 | 查全率 | F值 |
| 情感词 | 82.1% | 69.2% | 73.1% |
| 情感词+否定词+  程度词+标点符号 | 88.1% | 69.9% | 77.2% |
| 情感词+否定词+  程度词+标点符号  +n-gram | 91.2% | 71.5% | 80.2% |
| 情感词+否定词+  程度词+标点符号  +n-gram+词聚类特征 | 91.9% | 75.2% | 83.2% |

将实验数据用柱状图表示如下图 所示：

#### 图3-6 基于SVM的高维混合特征的弹幕文本情感倾向性分析折线图

从柱状图可以发现，本文改进的算法当特征组合更丰富的时候，查准率和查全率上都有着显著的提升。

2）结合客观语料的弹幕情感分类对比实验

此次试验准备基于上一章训练好的svm弹幕情感分类器模型为基础，为了研究模型预处理阶段是否考虑对客观语料做情感分析这一问题上，是否对最终的分析结果有所提升。

本文选取了5500条弹幕文本作为样本进行测试。其中标注主观语料3300条，情感客观语料2200条。

首先，结合客观弹幕的情感分析模型使用了通过模型生成的客观弹幕情感词典，具体生成的词典截图如下：

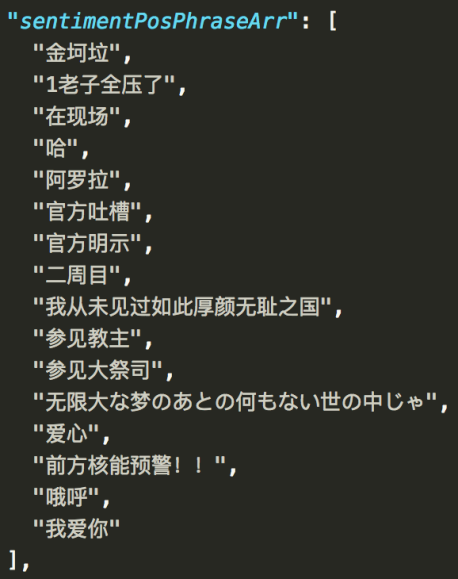
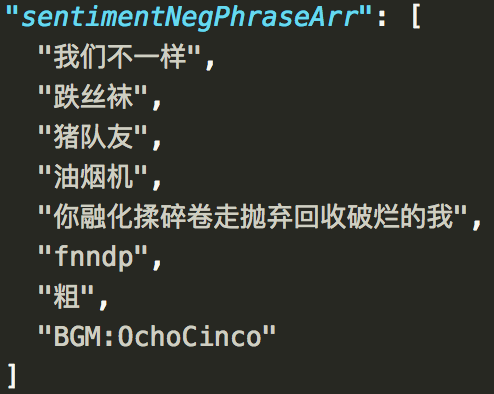


图3-7 客观弹幕词典截图（一）



#### 图3-8 客观弹幕词典截图（二）

其次，其实验结果如下表所示：

表4-9结合客观语料的弹幕情感分类对比实验数据表

|  |  |  |  |
| --- | --- | --- | --- |
| 模型选择 | 查准率 | 查全率 | F值 |
| 基础svm情感分类模型 | 91.9% | 75.2% | 83.2% |
| 结合客观语料的  基础svm情感分类模型 | 93.1% | 81.0% | 88.7% |

将实验数据用柱状图表示如下图 所示：

#### 图3-9 结合客观语料的弹幕情感分类对比实验折线图

## 3.4 本章小结

首先针对基于SVM的高维混合特征的弹幕情感分析算法做以小结。

本算法充分考虑弹幕短文本的特点，从多位混合特征的角度进行文本的特征抽取，做到尽可能兼顾语义和情感，并且取得了较好的实验效果，用实验的方法验证了本章方法的有效性和鲁棒性。本章算法主要的贡献有：提出了基于SVM的高维混合特征框架，采用了正则化的手段。考虑到语义对短文本情感分类的正确率影响，将词聚类加入到情感分析的特征，提高了一定的准确率。

第二部分算法研究针对于结合了客观弹幕语料的情感分析模型的研究。通过结合弹幕在时间轴序列上的密度排布，找出弹幕密集区域；通过主客观弹幕分类，主观语料输入分类器进行分类预测，客观弹幕进行高频短语提取，最终进行映射，构建一个客观情感弹幕词典。最终结合新生成的客观弹幕情感词典进行新的情感分类模型训练。在通过统一测试数据实验验证，结合客观弹幕的情感分类模型比普通的弹幕情感分类模型的预测准确度有明显的提高。

# 4 基于情感分析的视频弹幕系统需求分析

上一章主要介绍了本系统涉及到的基本概念与关键技术。从本章开始，将对该系统的框架结构、建模设计、实现测试进行详细介绍。在本章，主要从需求描述和分析模型两个角度进行论述。前者立足全局，将整个系统的框架和设计进行全面的概括；后者则将系统功能细分，详细介绍了系统需要实现的各大功能的具体设计。本章主要使用UML，从系统的功能模型，静态模型以及行为模型三个角度三个方面对需求设计进行规范化的描述。

## 4.1基于情感分析的视频弹幕系统需求描述

第二章中我们已经介绍过，弹幕文本会在一定程度上体现视频内容上的情绪波动情况。因此，分析视频弹幕文本的情感倾向，对分析视频的情绪倾向有重要的参考意义。然而，在当前大数据时代，信息量指数型增长，视频弹幕用户在海量的信息面前很难方便快捷的获取自己感兴趣的视频资源；并且现有的成熟视频弹幕站点上，也鲜有提供视频情节情感波动分析的功能。

基于情感分析的视频弹幕系统的基础还是网络视频，能够为用户提供完整的视频网站使用体验，可以便捷的搜索，播放视频，以及进行评论。同时，本系统提供了一些个性化的功能，用户可以在视频上发送弹幕，浏览弹幕；还可以看到当前时间下视频情节的情感波动情况。如上文所述，本文秉承个性化，实用性的原则，集合视频弹幕情感倾向分析分析功能，实现一个简单操作、可以实时发送弹幕，结果可视化，同时能为用户提供个性化的视频搜索浏览评论的在线视频平台。

本文的目标是建立一个基于情感分析的视频弹幕系统。用户在这里可以根据自己感兴趣的视频资源进行查询，了解最新最热门的网络视频资源。同时，本系统提供一定的个性化功能，用户可以查看自己的收藏的视频资源。此外，系统会针对视频资源拥有的弹幕文本进行情感倾向分析；因此，用户可以很方便的了解到当前视频的整体评价，方便帮助自己判断是否要观看当前视频。下面，本文将从系统的功能模型、静态模型和行为模型三个角度对系统进行需求分析。

## 4.2基于情感分析的视频弹幕系统的分析模型

本文设计的基于情感分析的视频弹幕系统将主要包括三种用户，即游客、会员以及管理员用户。下面的系统需求分析中，将先以用户的角度进行粗力度的功能划分，再按主要功能，结合用力图进行详细的功能模型轮式。

4.2.1基于情感分析的视频弹幕系统的功能模型

1）系统总体需求

本文的总体需求分析将从用户角色的角度进行描述。本文主要分为游客、会员和管理员三种角色。其中，游客可以浏览并查询系统中的视频资源，不仅是视频本身资源，还包括视频的相关信息等内容。会员需要登录系统，并且除了视频资源外，还可以进行发布弹幕文本，浏览弹幕等操作，也可使用视频论坛的相关功能。会员也可以对个人信息进行管理，包括密码、用户名等。将系统功能按照角色进行划分，可以得到如下图的系统游客和会员角色用例图：

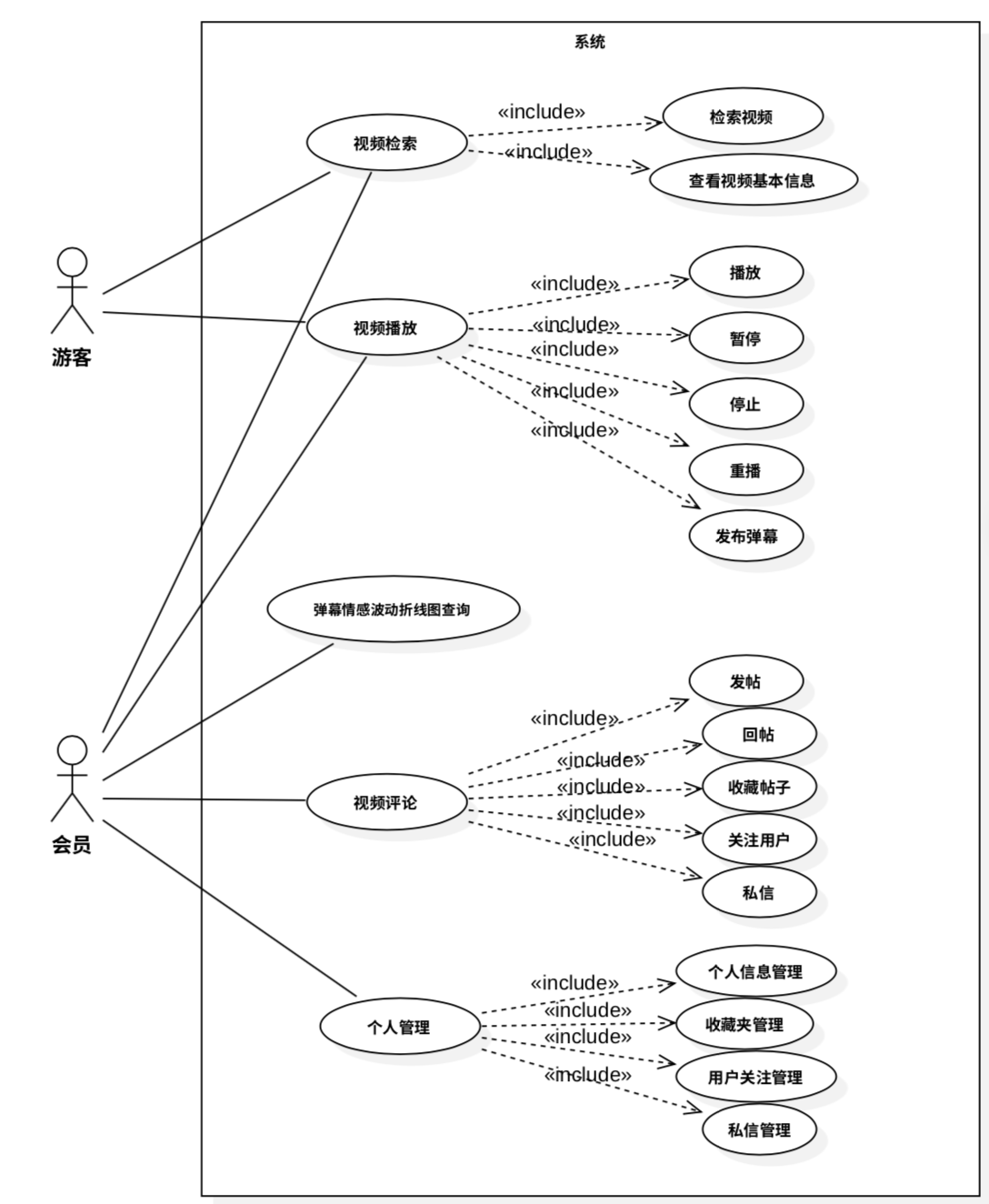


图4-1 游客与会员角色用例图

管理员是系统的管理者，可以对用户的信息、股票的交易数据、资讯数据、股票论坛的相关数据等进行管理。管理员角色的用例图如下图4-2所示：

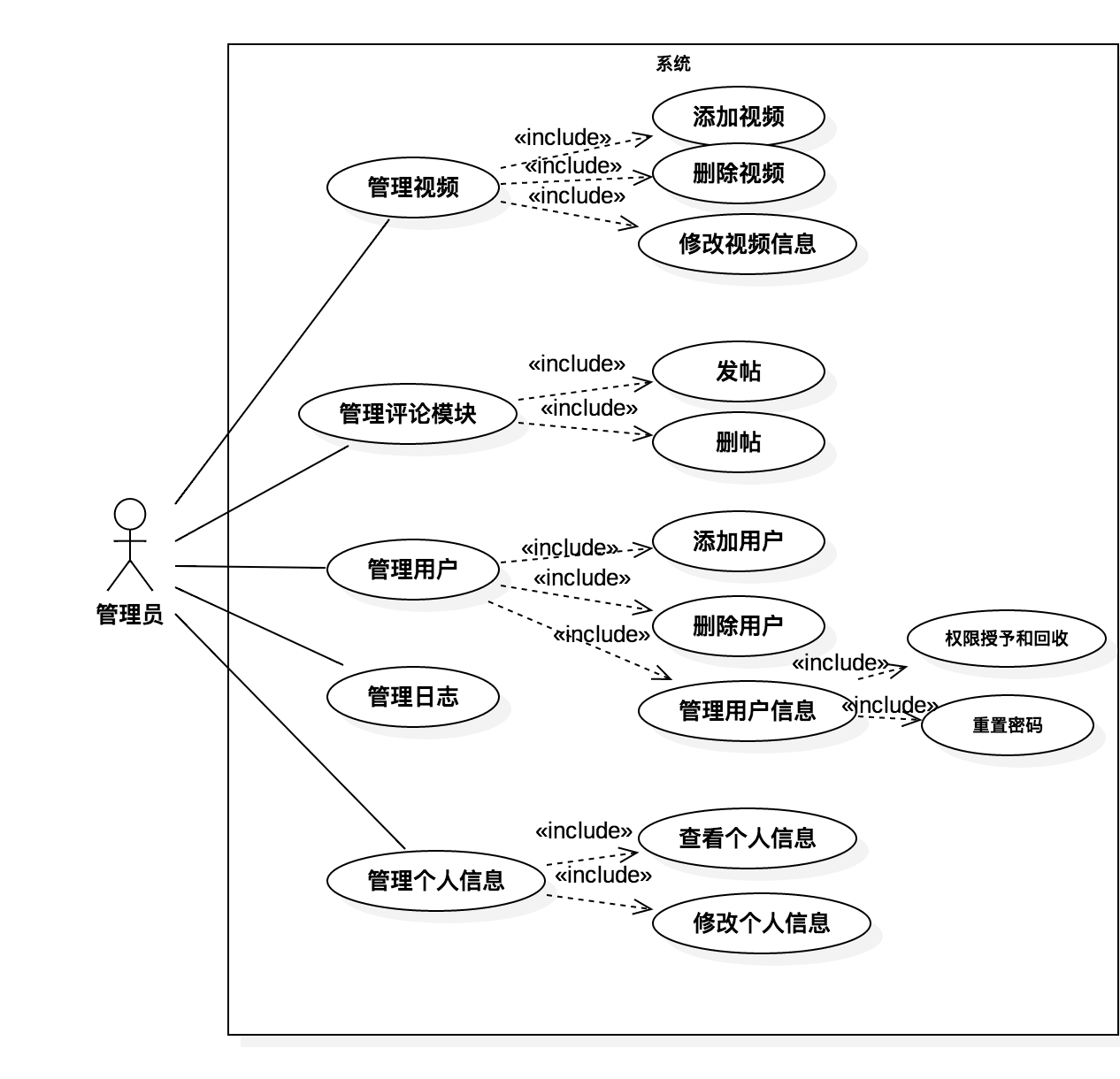


图4-2 管理员角色用例图

在对系统进行总体用例的分析之后，下面将就系统的具体功能需求，利用用例描述进行详细的分析。

2）视频资源查询

本系统的视频资源功能能够让用户方便的浏览感兴趣的视频资源和弹幕等相关数据，同时也能为用户提供个性化的视频弹幕情感倾向分析。

具体地说，视频查询功能可以分为视频检索、视频播放、视频弹幕查看、视频属性信息查看，视频评论查看五个用例。其中游客和会员都可以使用的功能是视频检索、视频播放、视频弹幕查看、视频属性信息查看，视频评论。视频检索为用户提供视频的查找功能，用户可以按照视频id、视频名称等关键字进行视频的信息检索。视频弹幕是在视频播放区域滑动显示的一种特殊的评论文本，它的特点是和视频时间有高度相关性。视频属性是视频资源的一些基本属性信息，包含了视频时长、视频来源、主题等。

下面以视频查询功能中的查看视频资源为例，对用例进行描述。查看视频资源的用例描述如下表4-1所示：

表4-1 视频查询用例描述

|  |  |
| --- | --- |
| 表项 | 说明 |
| 用例名称 | 查看视频资源 |
| 来源 | 需求 |
| 主要参与者 | 普通用户、会员用户 |
| 描述 | 该用例描述用户查询视频资源的过程 |
| 前置条件 | 该用户已经进入系统并登录 |
| 典型事件过程 | 1.用户进入系统首页  2.搜索框输入关键字并点击搜索  3.系统显示视频资源详情 |
| 后置条件 | 系统保存用户搜索记录 |
| 结论 | 当用户完成视频资源查询所有步骤，该用例结束 |

3）视频播放

用户通过查询用例找到需要的视频资源之后，点击进入播放界面。这一用例作为整个网站的核心功能，包含了用户控制播放视频资源的各项擦欧哦：播放，暂停，快进后退，全屏等操作。用例描述如下表4-2所示：

表4-2 视频播放用例描述

|  |  |
| --- | --- |
| 表项 | 说明 |
| 用例名称 | 播放视频资源 |
| 来源 | 需求 |
| 主要参与者 | 普通用户、会员用户 |
| 描述 | 该用例描述用户控制播放视频资源的过程 |
| 前置条件 | 该用户已经进入系统并登录 |
| 典型事件过程 | 1.用户进入视频资源页  2.视频默认自动播放  3.通过视频播放器内嵌的“播放”“暂停”“全屏显示”按钮控制视频播放，通过进度条拖拽控制视频播放进度 |
| 后置条件 | 系统保存用户播放记录 |
| 结论 | 当用户完成视频播放所有步骤，该用例结束 |

4）弹幕发布

弹幕是显示在视频表面浮层上的一种移动的评论文本，由于丰富的互动性备受视频网站用户的喜爱。本用例是提供用户发布弹幕的输入框，通过在特定的时间点键入弹幕文本，并点击发送按钮即可以发送弹幕。用例描述如下表4-3所示：

表4-3 弹幕发布用例描述

|  |  |
| --- | --- |
| 表项 | 说明 |
| 用例名称 | 发布弹幕文本 |
| 来源 | 需求 |
| 主要参与者 | 会员用户 |
| 描述 | 该用例描述用户发布弹幕文本的过程 |
| 前置条件 | 该用户已经进入系统并登录 |
| 典型事件过程 | 1.用户进入视频播放页  2.弹幕文本框输入并点击发布  3.弹幕区域出现当前弹幕 |
| 后置条件 | 系统增加一条弹幕记录 |
| 结论 | 当用户完成弹幕发布所有步骤，该用例结束 |

5）查看弹幕文本情感波动分析

通过第三章的弹幕文本情感分析算法分析，可以得到展示视频对应时间点上情感波动的折线图，这条曲线图是由一组时间序列上的情感值点经过连线生成的，当鼠标移动到某一个点上时，会弹出信息框，框中会显示对应的时间戳信息和情感值信息。用例描述如下表4-4所示：

表4-4 查看弹幕文本情感波动用例描述

|  |  |
| --- | --- |
| 表项 | 说明 |
| 用例名称 | 查看弹幕文本情感波动 |
| 来源 | 需求 |
| 主要参与者 | 会员用户 |
| 描述 | 该用例描述用户查看弹幕文本情感波动的过程 |
| 前置条件 | 该用户已经进入系统并登录 |
| 典型事件过程 | 1.用户进入视频资源播放页  2.系统显示弹幕文本情感波动详情  3.鼠标移动到折线的点上，弹出信息框显示对应属性信息 |
| 后置条件 | 系统保存用户浏览记录 |
| 结论 | 当用户完成查看弹幕文本情感波动所有步骤，该用例结束 |

6）视频评论模块

视频评论模块允许所有用户使用的系统功能。用户在登录系统后，点击视频资源链接，即可计入相应的视频资源页面，评论模块在视频资源页的下部。视频评论模块功能包括发帖、回帖、关注其他用户、收藏帖子以及向其他用户发送私信等。

下面以视频评论模块的发帖功能为例，对用例进行描述。视频评论模块发帖用例描述如下表4-5所示：

表4-5视频评论模块用例描述

|  |  |
| --- | --- |
| 表项 | 说明 |
| 用例名称 | 视频评论模块发帖 |
| 来源 | 需求 |
| 主要参与者 | 普通用户、会员用户 |
| 描述 | 该用例描述用户视频评论模块发帖的过程 |
| 前置条件 | 该用户已经进入系统并登录 |
| 典型事件过程 | 1. 用户进入系统首页 2. 搜索框输入关键字并点击搜索 3. 系统显示视频资源详情 |
| 后置条件 | 系统增加一条用户评论 |
| 结论 | 当用户完成发帖所有步骤，该用例结束 |

7）个人管理

会员在登陆系统后，可以进行个人管理的相关操作。用户可以进行个人信息的修改，也可以对收藏的视频、帖子、关注的用户等相关内容进行管理操作。除了和其他用户之间的私信操作，当用户对系统信息的准确性、系统检索的结果、或系统收录的信息的真实性有所怀疑时，都可以通过私信想管理员进行反馈。 下面以系统中的个人管理为例，对用例进行描述。个人管理的用例描述如下表4-6所示：

表4-6个人管理用例描述

|  |  |
| --- | --- |
| 表项 | 说明 |
| 用例名称 | 个人信息管理 |
| 来源 | 需求 |
| 主要参与者 | 普通用户 |
| 描述 | 该用例描述用户个人信息管理的过程 |
| 前置条件 | 该用户已经进入系统并登录 |
| 典型事件过程 | 1.用户点击“用户管理”按钮，进入用户管理界面  2.用户点击“基本信息”按钮，进行个人信息修改  3.用户对需要调整的信息进行修改或补充 |
| 后置条件 | 系统对用户最新的个人信息记录 |
| 结论 | 当用户完成个人信息管理的所有步骤，该用例结束 |

4.2.2基于情感分析的视频弹幕系统的静态模型

本文使用静态模型对系统进行分析的主要目标，是能够获得基于情感分析的视频弹幕文本系统的所有核心累，以及类之间的基本关系。系统的核心类图如下图4-3所示：

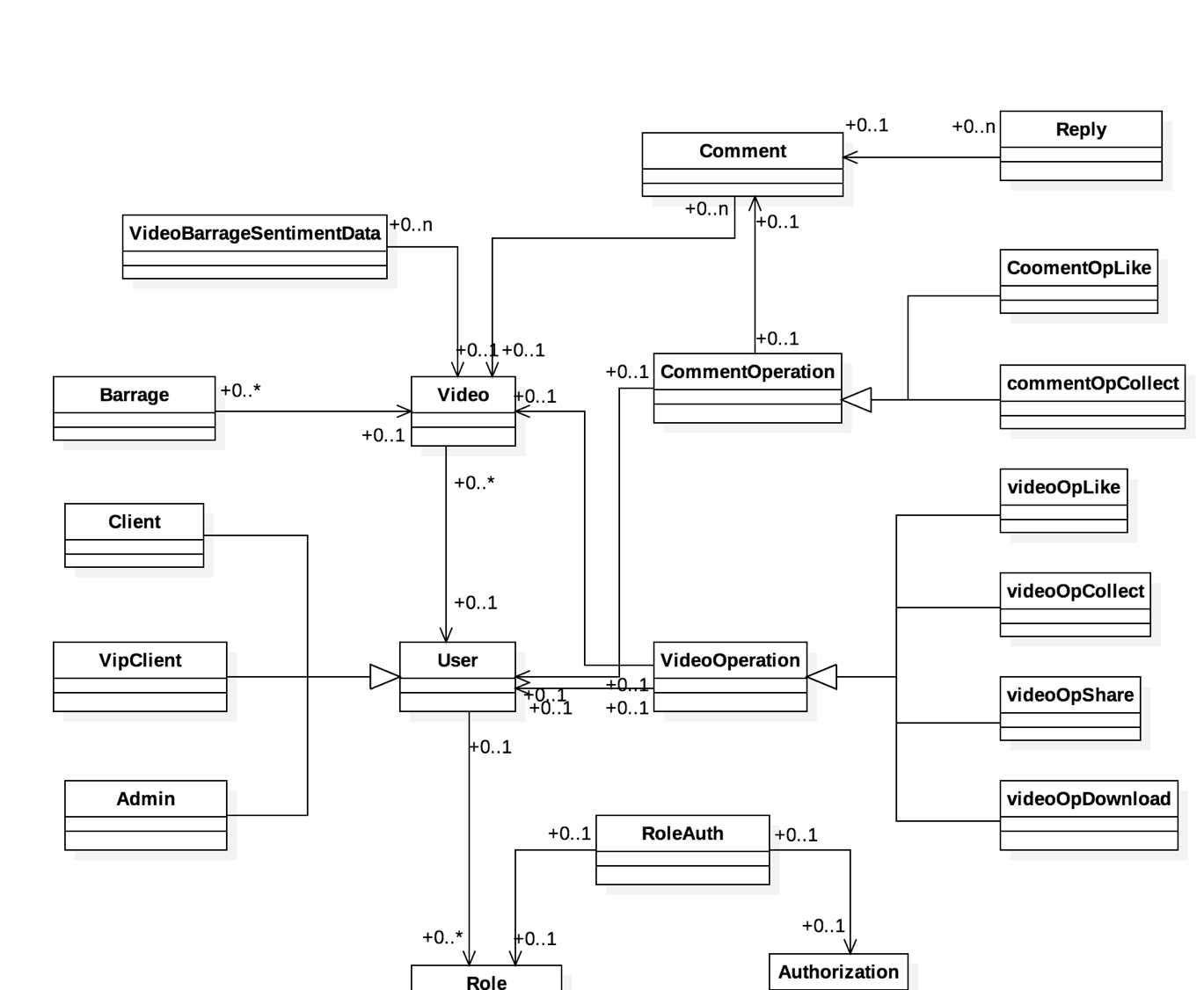


图4-3 系统类图

系统的核心类包括 等。其中，User类表示用户类，Video表示视频类，Barrage类表示弹幕文本类，这三个类是系统中关键的部分。

在对系统的静态模型进行说明之后，下面将对基于情感分析的视频弹幕文本系统的行为模型进行详细的描述。

4.2.3基于情感分析的视频弹幕系统的行为模型

本文将使用活动图来对基于情感分析的视频弹幕文本系统的行为模型进行说明。活动图用来描述业务用例内部时间六中不同活动之间的动作序列，实现不同互动之间的控制，特别适合于工作流和兵法的处理行为，同时可以按照需要来定义活动中的对象和对应的状态、角色和属性的改变。下面，将按照需求分析中的功能划分，对系统对象之间的交互过程进行详细的说明。

1）视频资源查询

系统的视频查询功能主要设计游客、会员用户和系统。下面以检索视频为例，说明用户对视频进行查询的过程：用户进入下同后，并在检索框中输入需要检索的额关键词。关键词可以是同意的视频编号、视频名称，也可以是视频发布者信息等。在输入检索关键词后，用户需要选择检索的信息类型。信息类型分为两类，一类是直接对视频进行检索，用户点击检索结果后就可以查看视频的技术指标等详细信息；另一类是该视频的相关附属信息。系统视频资源查询的活动图如下图4-4所示：

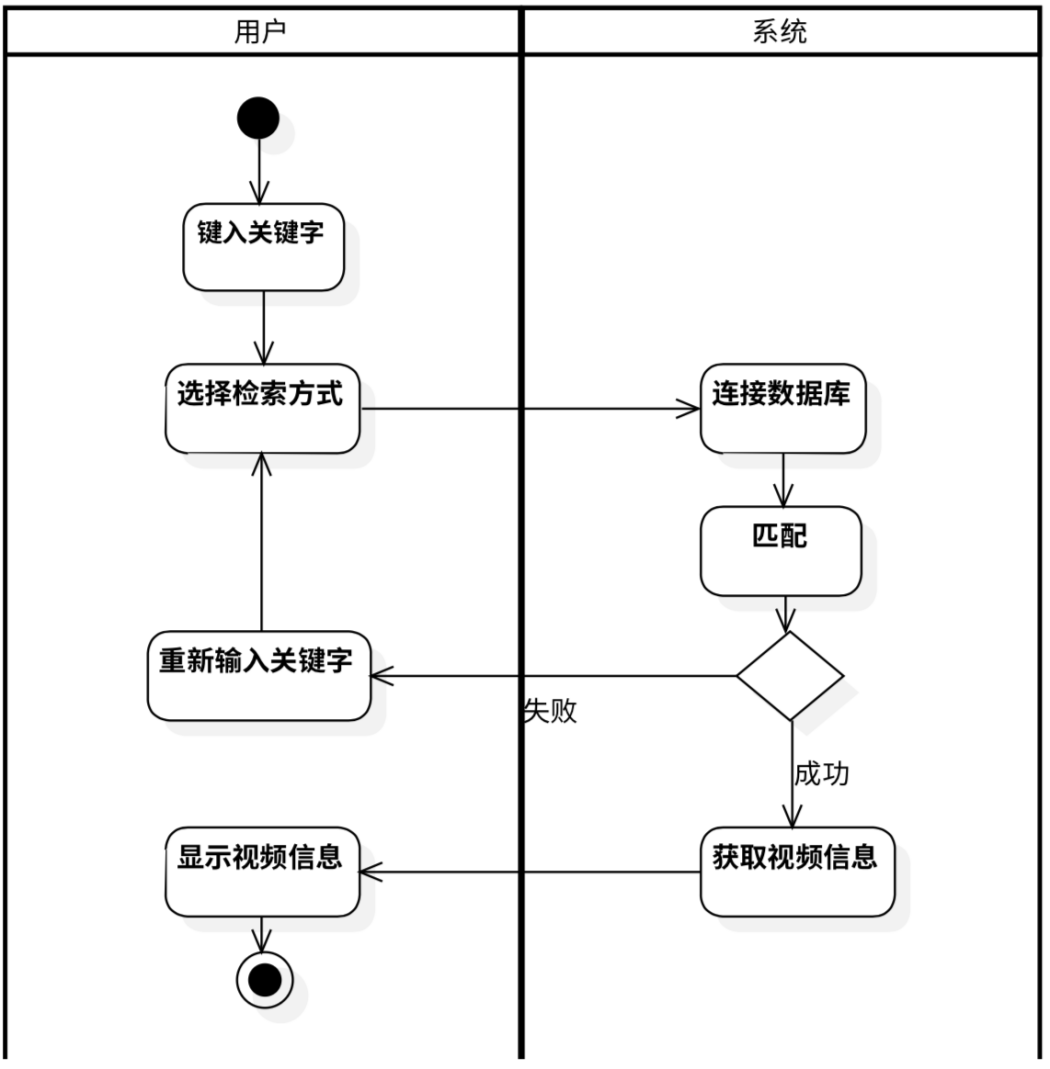


图4-4 视频资源查询活动图

2）播放视频

系统的基础功能是播放视频，并且系统提供的视频播放器提供了控制视频播放的一组按钮。用户首先登录系统，通过检索之后点击进入特定的视频播放页面。在页面中系统默认自动播放视频资源；用户可以通过控制条中的“暂停”、“停止”、“重播”按钮进行对视频的控制。

与此同时，在播放视频的同时，视频播放区域的浮层上会自动播放弹幕文本，用户可以观看弹幕文本；用户还可以在视频下部的弹幕发布输入框键入想发送的文本，点击发布按钮进行发布。播放视频的活动图如下图4-5所示：

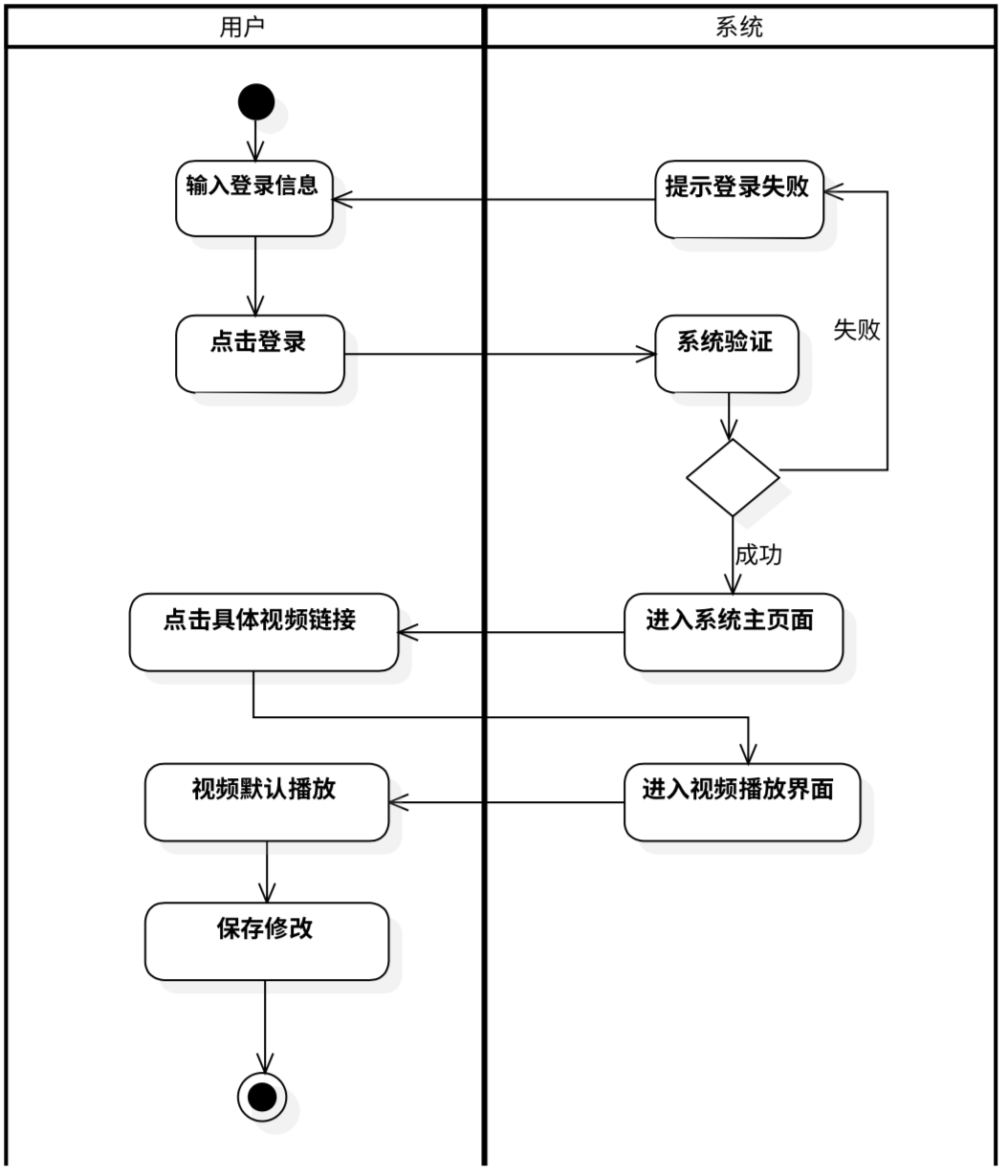


图4-5 播放视频活动图

3）查看视频情感波动折线图

用户可以在本系统中查看当前视频的情感倾向数值在时间轴上的波动状况，作为观看视频之前对当前视频简要了解的参考。用户需要首先在搜索框中输入视频编号或视频名称，在检索出需要查询的视频之后，进入视频详情页面。系统会对当前视频资源进行情感分析，生成一个横坐标是时间轴的情感波动折线图，显示在视频详情页中。查看视频弹幕情感波动折线图的活动图如下图4-6所示：

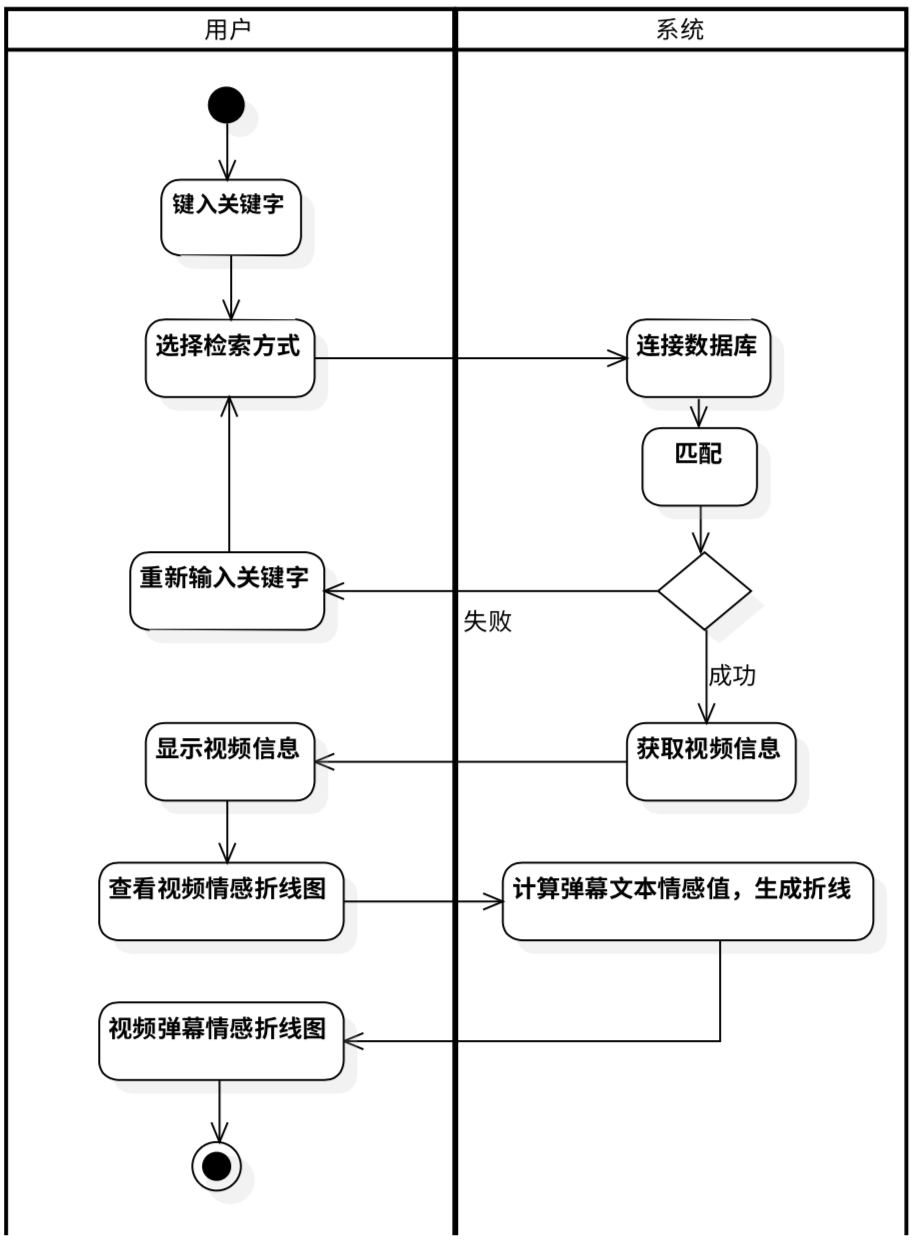


图4-6 查看视频情感波动折线图活动图

4）视频评论模块

视频评论模块的使用需要用户首先登录系统。用户可以在评论区发帖、回帖、关注用户、收藏帖子、收藏帖子或发送私信。用户视频评论模块的活动图如下图4-7所示：

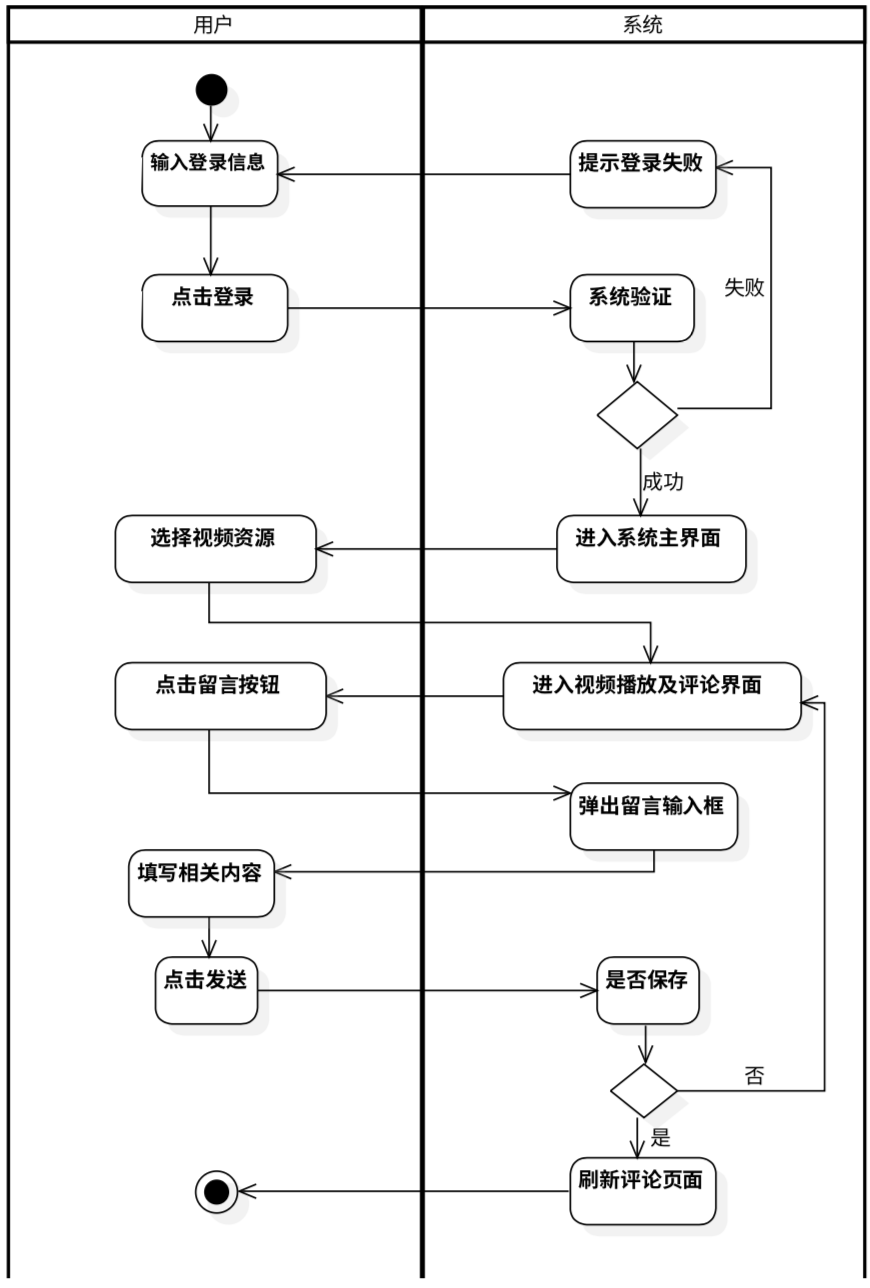


图4-7 视频评论模块活动图

5）个人管理

个人管理的本质是用户在登陆自己的账号后，对自身的信息，例如用户名、密码、收藏夹、私信等相关信息，进行增加、删除、修改、查找的操作。当用户正确登陆自己的账号之后，便可以进行相应的操作。现在以用户修改个人信息为例，介绍系统中用户进行个人信息管理的活动图，如图4-8所示：

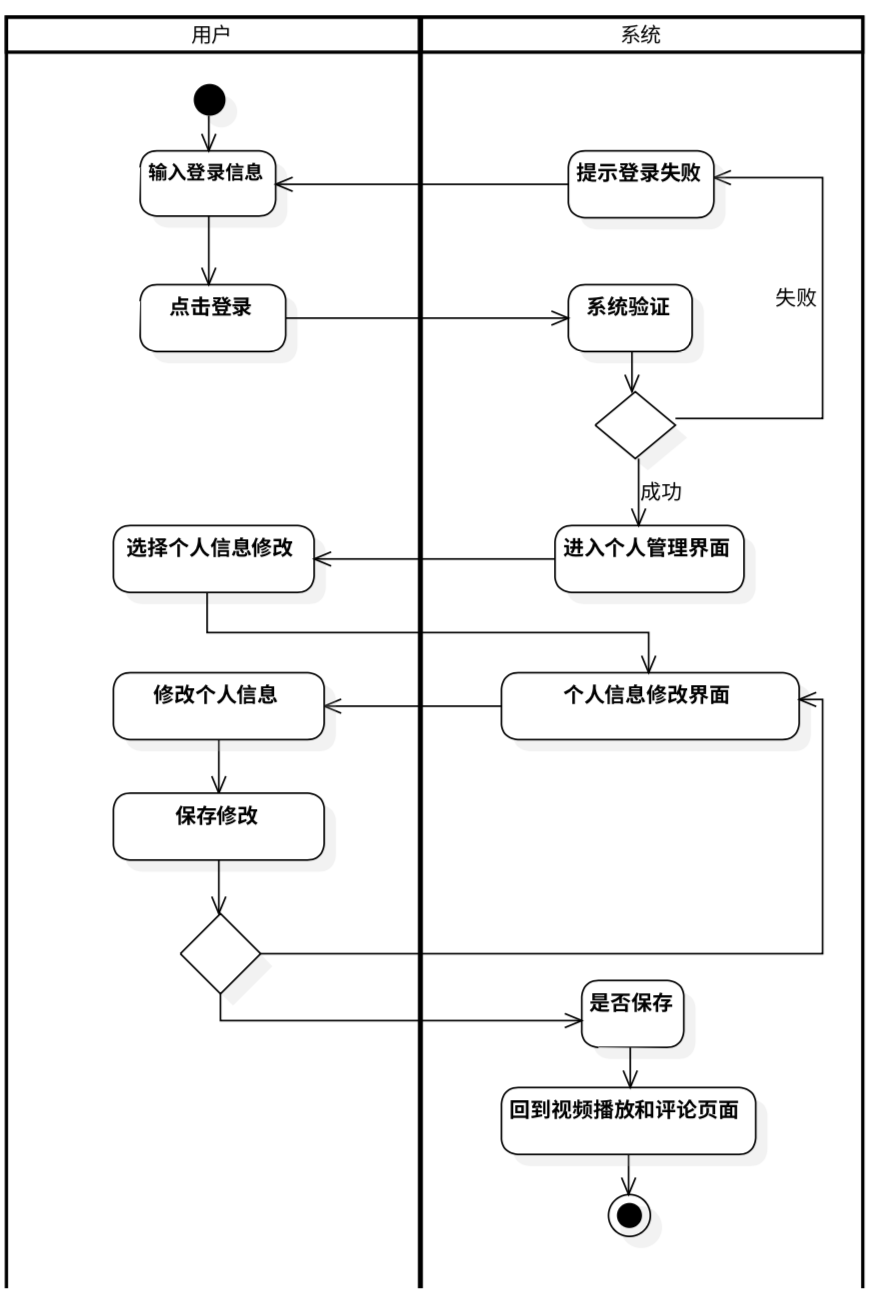


图4-8 个人信息管理活动图

## 4.3基于情感分析的视频弹幕系统的非功能性需求

系统的非功能性需求，是指产品为满足用户业务需求而具有的功能意外的其他特性。针对本文的基于倾向性的股票资讯服务系统，应当具有以下非功能性需求：

1）性能需求: 并发用户数小于1000，在1000个并发用户进行股票检索时，业务处理响应时间在5秒以内；对股票进行交易时，不计入网络传输时间，买进和卖出的响应时间在3秒以内。当使用股票论坛功能时，回帖、发帖的响应时间应该在5秒以内。

2）可靠性：可靠性是指，在规定的一段时间内与条件下软件维持其性能和水平的属性。软件的可靠性包括成熟性、容错性和易恢复性。尽管本文设计的交易行为是模拟、仿真的，但仍然应该对故障情况有严格的要求。此外，一旦发生故障，应当能够迅速重建系统，恢复数据，以免影响用户体验。

3）易用性：易用性是与一组规定或者潜在的用户为使用其软件所需做的努力和对这样的使用所作的评价有关的一组属性，包括易理解性、易学习性和易操作性。本文所设计的基于倾向性分析的股票资讯服务系统，在模拟交易方面应当具备和真实交易完全一致的使用体验，在股票舆情查询上则应该简便易学，便于用户迅速习惯。

4）可移植性：可移植性是指与软件可从某一环境转移到另一环境的能力有关的一组属性。具体包括系统的适应性、易安装性、遵循性、以及可替换性。软件的移动化是当前的趋势，股票的交易操作也越来越多得由手机完成，因此开发基于移动网络的版本对本系统有重要的意义。

5）安全性：即与防止对程序技术局的非授权的故意或者意外访问的能力有关的软件属性。如用户权限、动态口令、数据库字段加密等。系统应该能够保障用户的个人信息安全。

## 4.4 本章小结

本章对基于情感分析的视频弹幕系统的主要需求进行了分析和归纳，对系统需要实现的主要功能进行了清晰的表达和描述。本章首先立足全局，将整个系统的框架和设计进行了全面的概括：然后将系统功能进行了细分，详细介绍了系统需要实现的各大功能的具体设计。本章主要使用了UML，从系统的功能模型，静态模型以及行为模型三个角度三个方面对需求设计进行规范化的描述。

# 5基于情感分析的视频弹幕系统的设计

上章对基于情感分析的视频弹幕系统的整体需求做了介绍，对于系统需要完成的主要功能点进行了阐述，对主要角色的功能通过用例图进行了展示，同时进行了系统静态建模和系统动态建模。本章将从系统构架设计、系统概要设计、系统详细设计和数据库设计四个方面对系统设计进行讨论，为系统的实现打下基础。

## 5.1基于情感分析的视频弹幕系统的架构设计

本系统打算采用基于B/S（浏览器/服务器）模式的分层体系结构，表示层是返回给用户显示的页面，方便用户对系统的操作；数据层用来存放股票、用户信息等；业务层主要是事务处理，包括处理用户的请求、系统与用户的交互和实现系统的功能。 根据上一节对系统的需求分析，本文采用基于SSH（Structs+Spring+Hibernate）框架来构建系统，可以实现搭建结构清晰、可复用性好、维护方便的Web应用程序。其中使用Struts作为系统的整体基础架构，负责MVC的分离，在Struts框架的模型部分，控制业务跳转；利用Hibernate框架对持久层提供支持；Spring做管理，管理Struts和Hibernate。本系统的体系结构从职责上分为四层，依次为表示层、业务层、数据持久层和域模型层，如图5-1所示。

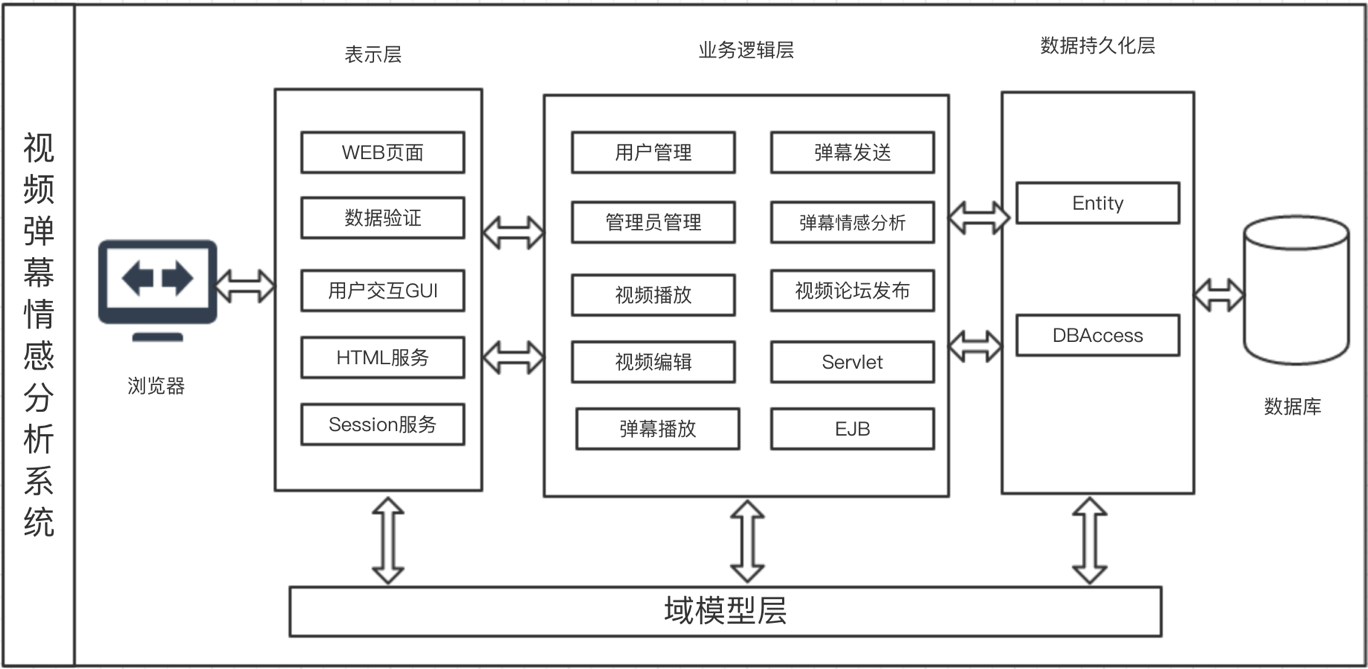


图5-1 系统的体系结构图

从图5-1中可以看出，表示层使用了Struts框架构建了系统页面与业务逻辑的分离，使页面显示和业务逻辑实现低耦合，业务层由Spring来完成事务的处理功能，数据持久层采用了Hibernate框架实现的DAO类来实现Java类与数据库之间的转换和访问，使业务逻辑与数据持久化分离。下面对各层功能做详细的论述。

表现层：包含系统与用户交互GUI（图形用户界面）和数据验证等。该层用于向系统用户提供图像交互界面，允许用户在显示页面中输入和编辑信息，同时该层还提供

数据验证功能，用于验证数据的有效性。该层主要包含的模块有系统各个功能模块的显示页面和相关请求处理页面。

业务层：包含业务规则的处理代码，即程序中与业务相关的专业算法、业务政策等。该层用于执行业务流程和指定数据的业务规则。业务层是面向业务应用，为表示层提供业务服务。主要包含的模块有视频播放、弹幕播放及发布、视频评论模块、情感分析可视化模块。还包括个人信息管理、用户管理、帖子管理、评论管理模块中需要进行处理的部分，需要与数据库之间交互处理。

数据持久层：包含数据处理代码和数据存储代码。该层主要包括数据存取服务，负责与数据库之间的通信。包括视频、资讯、评论、账户信息、弹幕信息、用户和管理员的个人信息。

## 5.2基于情感分析的视频弹幕系统的概要设计

5.2.1 功能模块设计

本文设计的基于情感分析的视频弹幕系统的最终目的，是要实现用户方便查询及播放视频，观看弹幕，并结合弹幕的情感分析可视化帮助用户更好的观看视频的需求。用户在该系统中可以获得当前视频情感值在时间轴上的波动情况，通过折线图的波动情况告诉用户哪些时间点上是视频情节冲突的高潮，从而帮助自己决策观看视频。 从系统的功能角度，可以将本系统具体的分为5个功能模块，分别是视频资源查询模块、视频播放模块、视频评论模块、个人管理模块、管理员管理模块。下面对系统的各个功能模块进行详细的介绍。系统功能模块图如下图5-2。

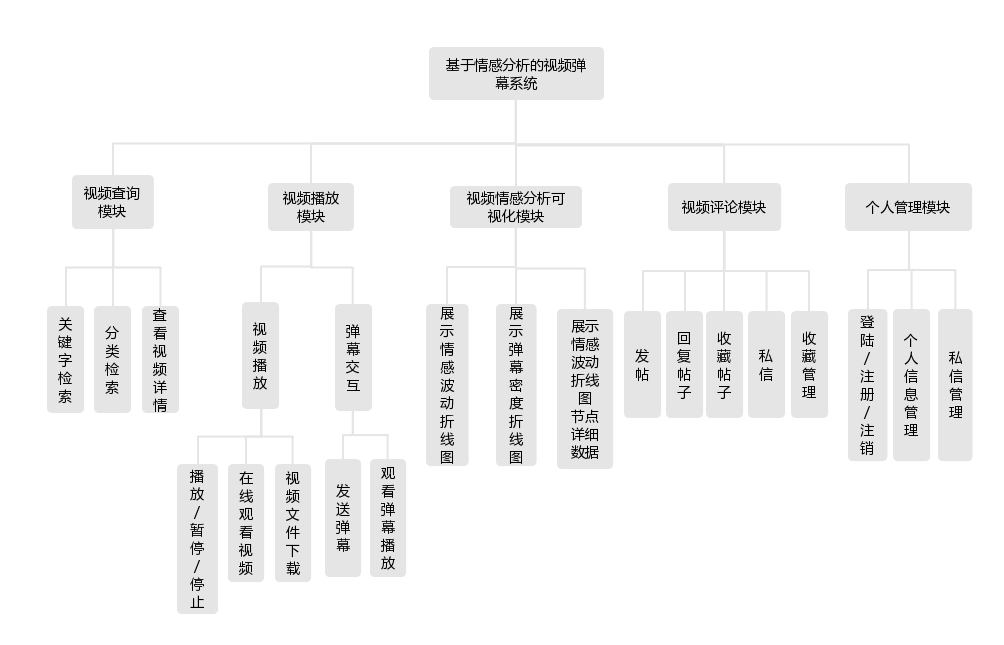


图5-2 系统的功能模块图

视频资源查询模块，是用户进行视频资源浏览的功能。普通游客用户可以对具体的视频资源进行检索，以检索自己感兴趣的视频或者相关的视频资源。游客也可以浏览系统视频首页的视频资源列表，对系统默认显示的视频信息进行查看。用户登陆系统并认证为会员后，就可以使用系统资讯查询的个性化功能。会员可以选择查看的视频的弹幕，也可以查看股视频弹幕分析生成的情感可视化图表，从而掌握视频的情感波动情况。

视频播放模块，是整个基于情感分析的视频弹幕系统的核心模块，其中由三个子模块构成。其一是视频播放模块，用户通过视频播放页面中的视频播放器，可以观看和控制视频，如暂停、停止、重播等。其二是视频弹幕子模块，功能集成了弹幕文本的播放，以及用户发布弹幕的输入框。用户可以在此键入自己的弹幕文本并发送。其三是视频弹幕情感分析可视化子模块，通过系统对弹幕的情感分析，会借助可视化的形式在页面对应区域显示分析的结果，包括了视频密度折线图、视频弹幕情感波动折线图等。

视频论坛模块，会员在登陆系统后，可以使用视频评论区中的相关功能。用户可以在视频评论区中交流自己的观看体验，也可以发表对视频呢的看法等等。本模块的功能包括发帖、回帖、关注其他用户、收藏帖子以及向其他帖子发送私信。该模块为系统增加了社交属性，增强了用户黏性，为用户的交流、沟通提供了平台。

个人管理模块，是指会员用户对个人信息等相关信息进行管理的功能。系统使用者首先可以使用用户注册的功能，成为系统的注册用户，即会员。在注册成为会员后，用户就可以进行登陆操作，并使用会员的相关功能。用户在登陆后，可以进入个人信息管理界面，对账号信息，例如用户名、密码、头像等进行修改。用户在个人管理模块下，也可以进行收藏的管理，浏览、删除已经收藏的股票或帖子。用户同时可以对关注的其他用户进行删除。此外，在个人管理功能中，用户可以对私信进行查看、删除等操作。

管理员管理模块：管理员管理是针对系统管理者的，系统管理员不需要注册，他们的账号是在系统后台中直接分配的。当系统管理员管理系统的时候首先需要以管理员的身份登录，登录以后可以管理个人资料、可以对系统用户进行管理，包括系统用户的增加和删除等。系统管理员还可以对用户的评论和反馈进行管理，可以通过查看使用系统时产生的日志管理查看系统是否正常运行。当然，管理员也可以对股票的信息进行管理，并就论坛的内容进行添加、搜索、修改、删除的操作。当管理员退出系统时需要注销系统，保证系统的安全。

5.2.2 类的设计

系统的详细类图如下图5-3所示：

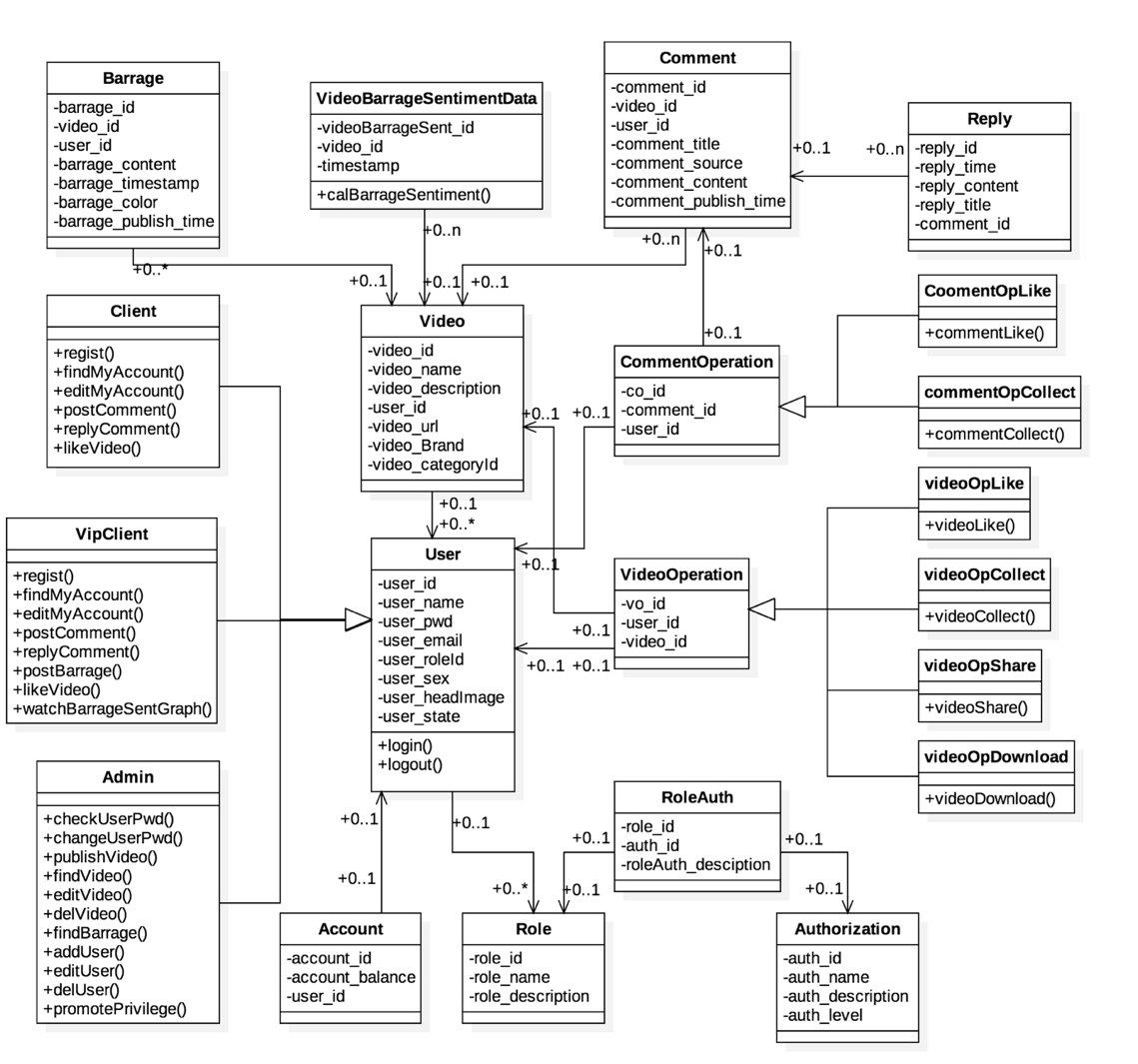


图5-3 系统的详细类图

图中可以看到，系统中最核心的类是User类，Video类，Barrage类以及SentiAnalyze类。

User类涉及login()和logout()方法，分别用于系统的登录和注销。Client、VipClient和Admin类继承自User类。其中Client类是系统的普通用户类，它包含有一些用户通用的方法，如regist()注册用户,editMyAccount()编辑个人帐户信息,postComment()发布评论,replyComment()回复别人的评论,likeVideo()给视频点赞等等。VipClient类是系统的会员类，所设计的方法有postBarrage(),watchBarrageSentiGraph()等，分别指的是发布弹幕、查看弹幕情感分析折线图，它们是会员特有的功能方法。

Video类是系统的视频资源类，它有较为详细的属性信息，相对完整的描述了视频的基本状况。用于展示从主流视频网站上爬取的视频资源，这些视频网站主要是中国最大的弹幕视频网站bilibili.com。

Barrage类代表的是弹幕类，与视频类Video、用户类VipClient等关联，支撑视频弹幕的功能业务，对用户对视频的弹幕评论进行表示。

在情感分析类SentiAnalyze中，calBarrageSentiment()是情感分析所需要的方法，并记录各视频计算得出的情感分类结果信息。它与视频类Video、用户类VipClient等关联，对视频弹幕的情感走势进行了分析，最终对分类结果进行统计呈现给用户。

在对系统的类进行描述之后，将利用类中的方法，进一步对系统的工呢个进行详细设计。

## 5.3基于情感分析的视频弹幕系统的详细设计

本节对系统的详细设计的阐述将类的设计以及各功能模块的详细设计方面展开。在此前对于系统功能分析的基础上，本节将对系统功能进行进一步的详细设计，主要分析系统的5个子功能模块，并给出各模块的顺序图。

通过前面对系统功能的分析，本节对系统功能的详细设计主要从视频资源查询模块、视频播放模块、弹幕情感分析模块、视频评论模块、个人管理模块等给予详细的阐述，并给出各功能模块的顺序图。

5.3.1 视频查询模块

视频检索是系统查询模块的子功能。用户进入系统主页，选择搜索方式为搜索视频，在搜索框中输入查询关键词，比如视频名称或者id，系统对输入的关键词进行匹配，数据库最终返回检索结果给用户。用户检索视频的时序图如图5-4所示。

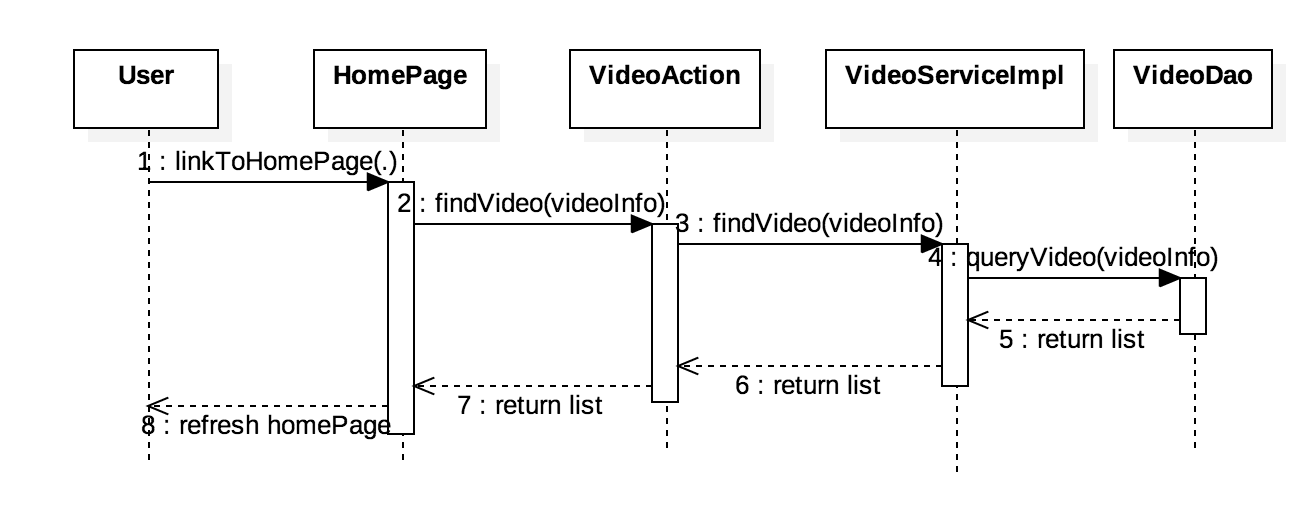
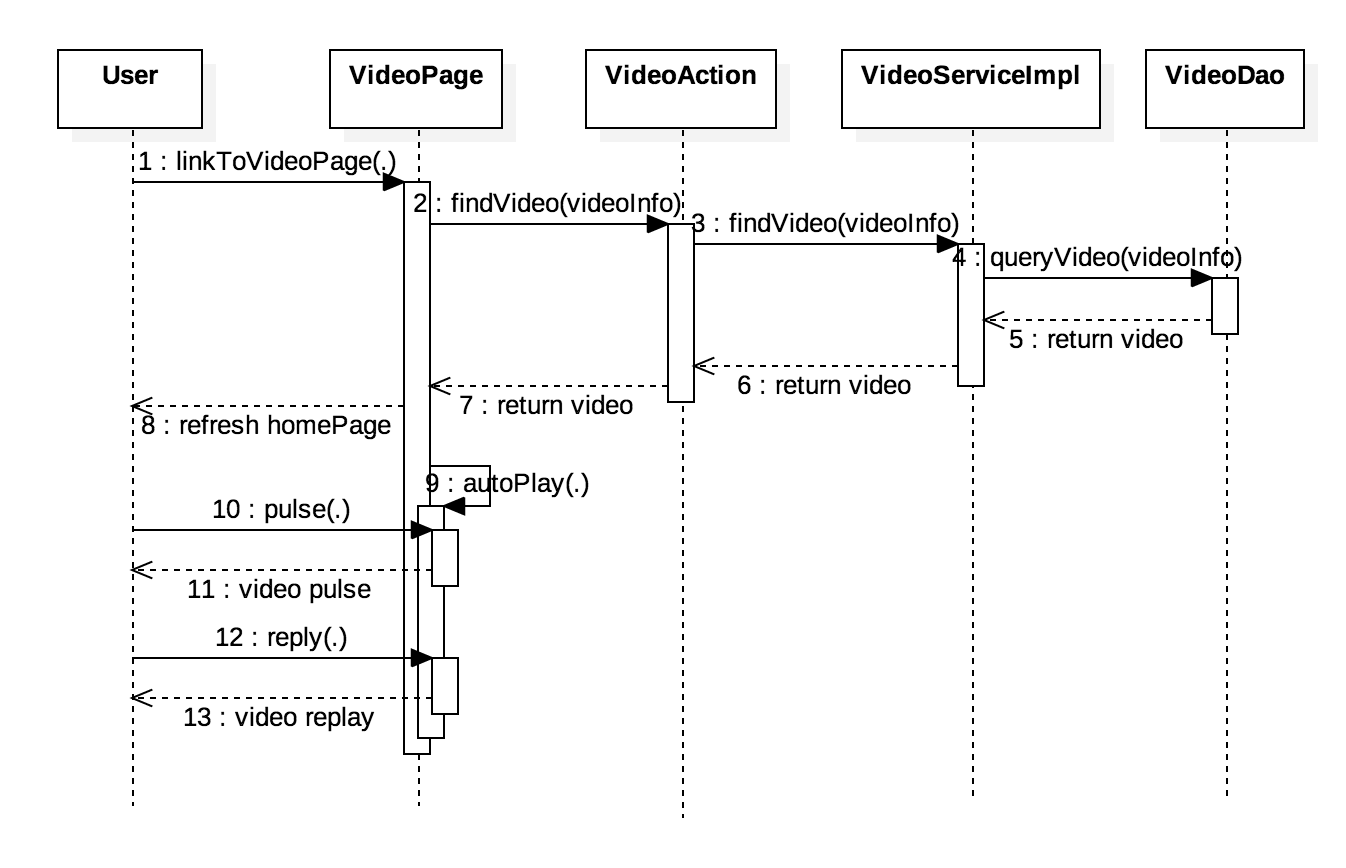


图5-4 用户视频查询时序图

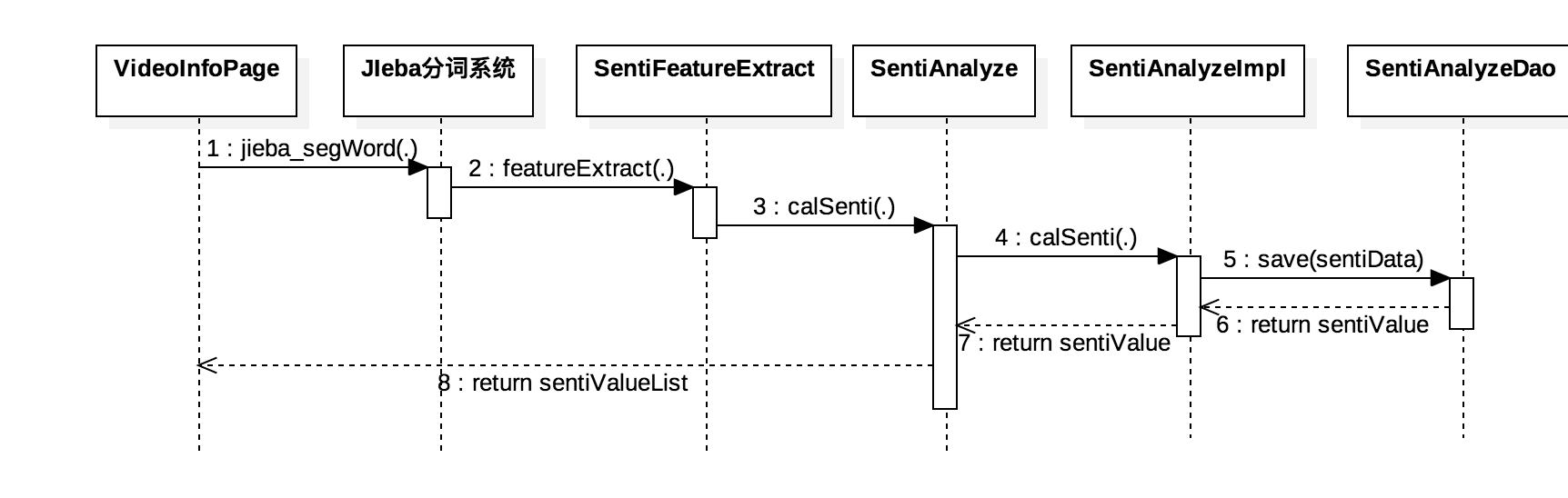
5.3.2 视频播放模块

视频播放模块，顾名思义是指对视频文件进行播放的模块，也是整个视频弹幕网站的基础模块。页面在刷新时会自动播放视频文件；在观看视频的同时，用户可以通过视频播放器完成对视频的控制。系统用户视频播放模块时序图如图5-6所示。

图5-5 视频播放模块时序图

### 5.3.3 弹幕情感分析模块

视频情感分析是系统针对视频的弹幕文本进行的情感分析的过程。系统首先对弹幕文本进行分词等预处理操作；随后进行词性情感特征抽取；之后根据本文改进的结合客观弹幕的情感分析模型进行情感模型优化，并结合SVM算法最终得出弹幕文本的情感分类，统计所有相关弹幕文本的情感分类，将结果返回给视频展示页面，用于展示视频的情感波动的可视化数据。系统视频情感分析的时序图如图5-5所示。

图5-6 弹幕情感分析模块时序图

5.3.4 视频评论模块

视频评论模块，用户可以发帖和回复。用户在评论区发帖的顺序图如图5-7所示。

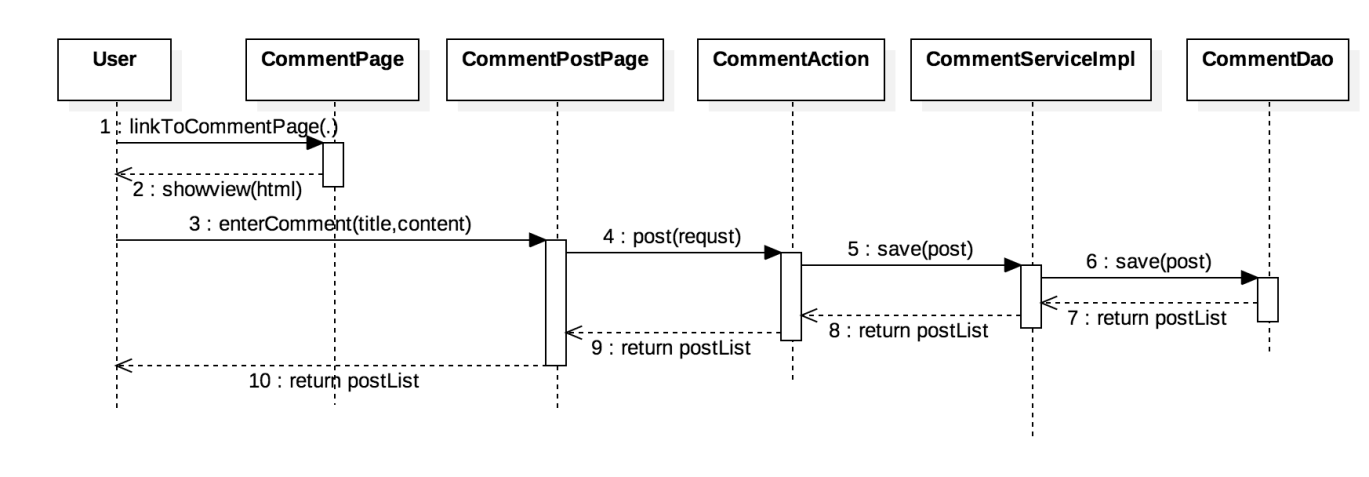
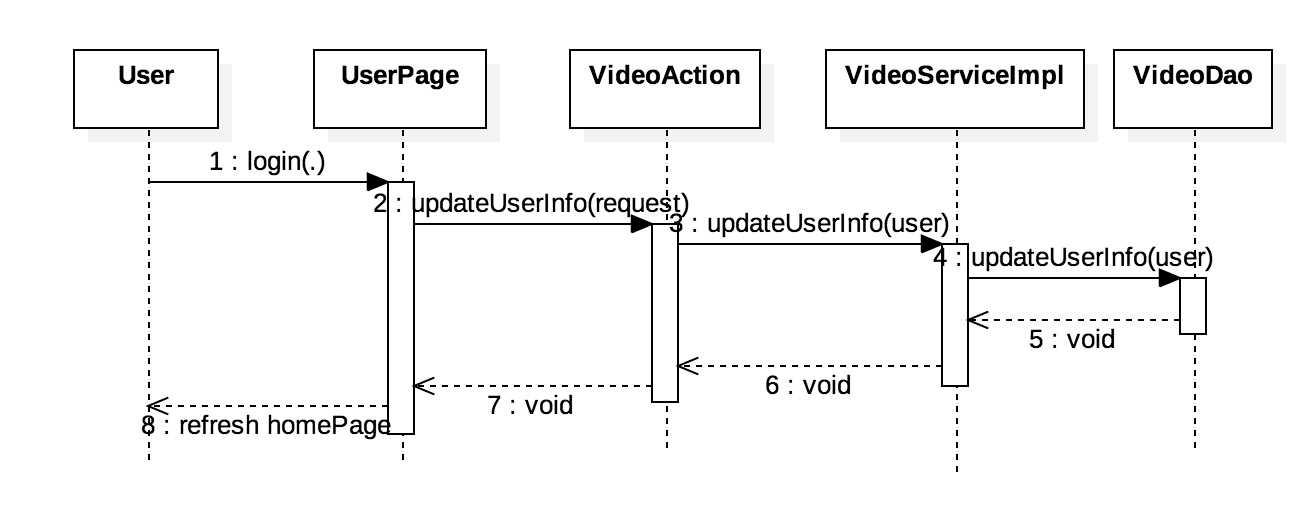


图5-7 用户评论时序图

5.3.5 个人管理模块

在个人管理模块中，用户可以对自己的个人信息进行修改，包括登录密码、邮箱、头像等账号信息，以及收藏信息，关注用户等。个人信息修改的时序图如图5-8所示。

图5-8 个人信息修改时序图

## 5.4基于情感分析的视频弹幕系统的数据库设计

本文使用Mysql数据库作为基于情感分析的视频弹幕系统的数据库。系统的数据库设计如下图5-10所示：

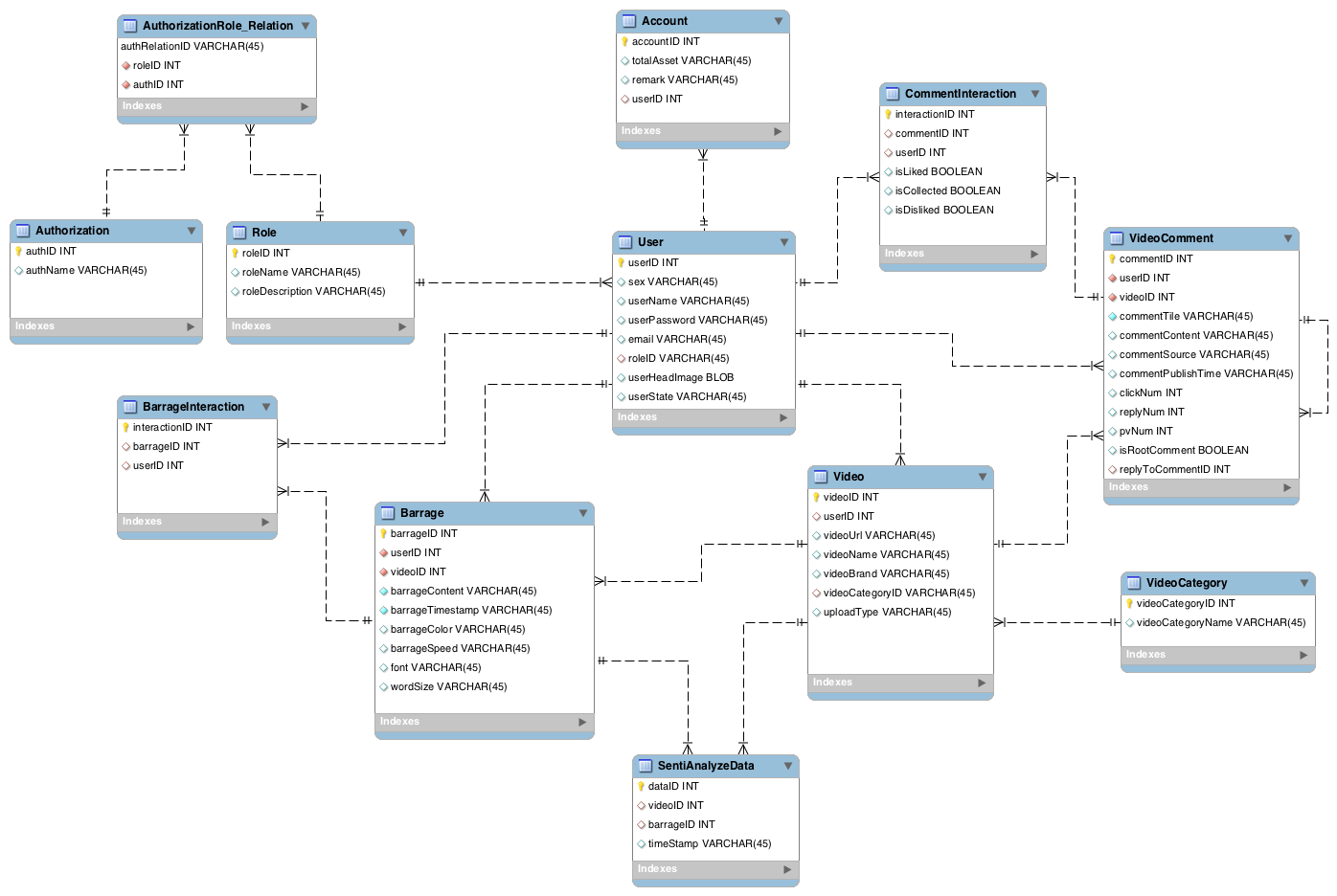


图5-9 系统数据库的实体关系模型

数据库业务围绕User表，Video表，Barrage表和VideoComment表进行设计。整个系统的业务，均由User表发起。User表根据角色类型分为管理员和普通用户、VIP用户。Video表存储视频资源，支持系统的视频业务。Barrage表存储弹幕基本信息，方便和其他表单进行关联。VideoComment表存储和视频评论、回复等信息，以及它们之间的依存关系。

同时，有一些表用来存储用户的一些交互信息。比如CommentInteraction表存储了用户对已存在的评论信息的一些交互操作，如点赞、踩、收藏等；又如BarrageInteraction表存储了用户对已发布的弹幕的交互操作，如点赞等。

Authorization表和Role表分别表示权限表和用户角色表，通过AuthorizationRole\_Relation表关联起来。以上三个表主要支撑系统的用户权限配置模块的持久层功能。

有了以上的分析，下面可以对数据库的各个表格，及表格中字段和类型进行具体的设计。由于篇幅的原因，本文就系统最重要的几张数据库表进行介绍，包括用户表、视频表、弹幕表、视频评论表、角色表、权限表，各个表的结构如下 。

1）用户表

用户信息表记录用户的基本信息，其中的UserID代表用户编号，也是用户信息表的主键，UserName代表用户名，UserPwd代表用户密码， RoleID代表用户的角色编号，用户角色一共有两种，分别为普通用户和管理员用户。用户信息表如表5-1所示。

表5-1 用户信息表

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段代码 | 数据类型 |
| 用户ID | UserID | Integer |
| 角色ID | RoleID | Integer |
| 用户名 | UserName | Varchar(45) |
| 用户密码 | UserPwd | Varchar(45) |
| 用户邮箱 | UserEmail | Varchar(45) |
| 用户头像 | UserHeadImage | Varchar(45) |
| 用户状态 | UserState | Varchar(45) |
| 用户性别 | UserSex | boolean |

2）视频表

视频表主要有如下关键字段：视频资源路径描述视频资源的文件位置，视频分类ID表示了视频所属的戏份类别的ID。详细字段介绍如表5-2。

表5-2 视频表

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段代码 | 数据类型 |
| 视频ID | videoID | Integer |
| 用户ID（外键） | userID | Integer |
| 视频资源路径 | videoUrl | Varchar(45) |
| 视频名称 | videoName | Varchar(45) |
| 视频品牌 | videoBrand | Varchar(45) |
| 视频分类ID | videoCategoryID | Integer |

3）弹幕表

弹幕表主要有如下关键字段：弹幕时间戳指向弹幕所对应的视频播放时间点。详细字段介绍如表5-3。

表5-3 弹幕表

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段代码 | 数据类型 |
| 弹幕ID | barrageID | Integer |
| 用户ID（外键） | userID | Integer |
| 视频ID（外键） | videoID | Integer |
| 弹幕文本内容 | barrageContent | Varchar(45) |
| 弹幕时间戳 | barrageTimestamp | Timestamp(10) |
| 弹幕颜色 | barrageColor | Varchar(45) |
| 弹幕字体 | Font | Varchar(45) |
| 弹幕字号 | wordSize | Varchar(45) |

4）视频评论表

视频评论表主要有如下关键字段：评论来源这一字段用来区分是本系统发布的评论或者来自外部论坛爬去的评论数据；是否根评论字段，区分是根评论还是对其他评论的回复；回复对象评论ID即如字面意。详细字段介绍如表5-4。

表5-4 视频评论表

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段代码 | 数据类型 |
| 评论ID | commentID | Integer |
| 用户ID（外键） | userID | Integer |
| 视频ID（外键） | videoID | Integer |
| 评论标题 | commentTitle | Varchar(45) |
| 评论内容 | commentContent | Varchar(45) |
| 评论来源 | commentSource | Varchar(45) |
| 评论发布时间 | commentPublishTime | Timestamp(10) |
| 点击数 | clickNum | Integer |
| 回复数 | replyNum | Integer |
| 页面浏览数 | pvNum | Integer |
| 是否根评论 | isRootComment | Boolean |
| 回复对象评论ID | replyToCommentID | Integer |

5）角色表

角色表详细字段介绍如表5-5。

表5-5 角色表

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段代码 | 数据类型 |
| 角色ID | roleID | Integer |
| 角色名称 | roleName | Varchar(45) |
| 角色描述 | roleDescription | Varchar(45) |

6）权限表

权限表详细字段介绍如表5-6。

表5-6 权限表

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段代码 | 数据类型 |
| 权限ID | authID | Integer |
| 权限名称 | authName | Varchar(45) |
| 权限描述 | authDescription | Varchar(45) |
| 权限粒度级别 | authLevel | Varchar(45) |

## 5.5 本章小结

本章根据上一章的需求分析和功能设计，对该系统需要使用到的主要模块进行了详细的分析和设计。在整体框架的指导下，分别对系统的视频资源查询模块、视频播放模块、视频评论模块、个人管理模块、管理员管理模块进行了详细的刻画。借助UML工具，将系统的主要内容进行了设计，包括架构设计、概要设计、详细设计和数据库设计等。在整个设计的过程中，该系统遵循系统分析的要求，对系统的具体功能进行了完整的覆盖，并在弹幕情感分析模块对关键的算法进行了改良，使得系统的视频弹幕情感分析更加优化，并且具有更好的可用性。本章为下一章的系统实现奠定了基础。

# 6基于情感分析的视频弹幕系统的实现与测试

## 6.1 系统的开发环境

6.1.1 系统的软件配置

操作系统：Windows864bit

数据库：MySQL5.6.16

JAVA开发环境：MyEclipse8.5，JDK1.6.0\_17

服务器：Tomcat6.0

6.1.2 硬件平台参数

设备：ThinkPadT420

CPU：酷睿i52520M

内存：4GB

硬盘：500GB

## 6.2 系统主要功能的实现

6.2.1 视频检索模块

1）关键代码

@RequestMapping( value = "videoSearchInit" )

public ModelAndView videoSearchInit( @RequestParam("videoCategoryID") int videoCategoryID, @RequestParam("page") int page, HttpServletRequest request )

{

ModelAndView mv = new ModelAndView( "VideoSearch" );

List<VideoCategory> videoCategoryList = vm.getVideoCategoryList();

List<Video> videoList = vm.getVideoListByVideoCategroyIDAndPage( videoCategoryID, page );

int videoListSize = vm.getVideoListSizeByVideoCategoryID( videoCategoryID );

if ( videoCategoryList != null && videoList != null )

{

mv.addObject( "videoCategoryList", videoCategoryList ); /\*传给视频分类栏使用 \*/

mv.addObject( "videoList", videoList ); /\*传16个视频对象给16个视频区域使用 \*/

mv.addObject( "videoListSize", videoListSize ); /\*传给分页组件使用 \*/

mv.addObject( "videoCategoryID", videoCategoryID );

mv.addObject( "page", page );

}

return(mv);

}

@RequestMapping(value="ajax\_search\_videoConsole\_videoPlayInfo", method=RequestMethod.GET, produces = "text/html;charset=UTF-8")

@ResponseBody

public String ajax\_search\_videoConsole\_videoPlayInfo(String videoId, String videoPlayDatetimeStart, String videoPlayDatetimeEnd, HttpServletResponse response) throws JsonProcessingException, ParseException{

HttpSession session = httpServletRequest.getSession();

TblUser user = (TblUser) session.getAttribute("user");

HashMap<String , Integer> videoPlayInfo = new HashMap<String , Integer>();

TreeMap<String , Integer> playCountList = new TreeMap<String , Integer>( SimpleDateFormat df = new SimpleDateFormat("yyyy-MM-dd");

java.util.Date videoPlayDatetimeStart\_temp1 = df.parse(videoPlayDatetimeStart);

java.sql.Date videoPlayDatetimeStart\_final = new java.sql.Date(videoPlayDatetimeStart\_temp1.getTime());

java.util.Date videoPlayDatetimeEnd\_temp1 = df.parse(videoPlayDatetimeEnd);

java.sql.Date videoPlayDatetimeEnd\_final = new java.sql.Date(videoPlayDatetimeEnd\_temp1.getTime());//为了将2017-01格式转换成2017-01-01的格式的类型

videoPlayInfo = vm.getPlayInfo(videoId, videoPlayDatetimeStart\_final, videoPlayDatetimeEnd\_final);//获取当前视频的播放统计数据

playCountList = vm.getPlayCountList(videoId, videoPlayDatetimeStart\_final, videoPlayDatetimeEnd\_final);//获取当前视频的播放统计折线图数据

List<Object> returnObj = new ArrayList<Object>();

returnObj.add(videoPlayInfo);

returnObj.add(playCountList)

ObjectMapper mapper = new ObjectMapper();

String jsonString = mapper.writeValueAsString(returnObj);

return jsonString;

}

2）实现效果

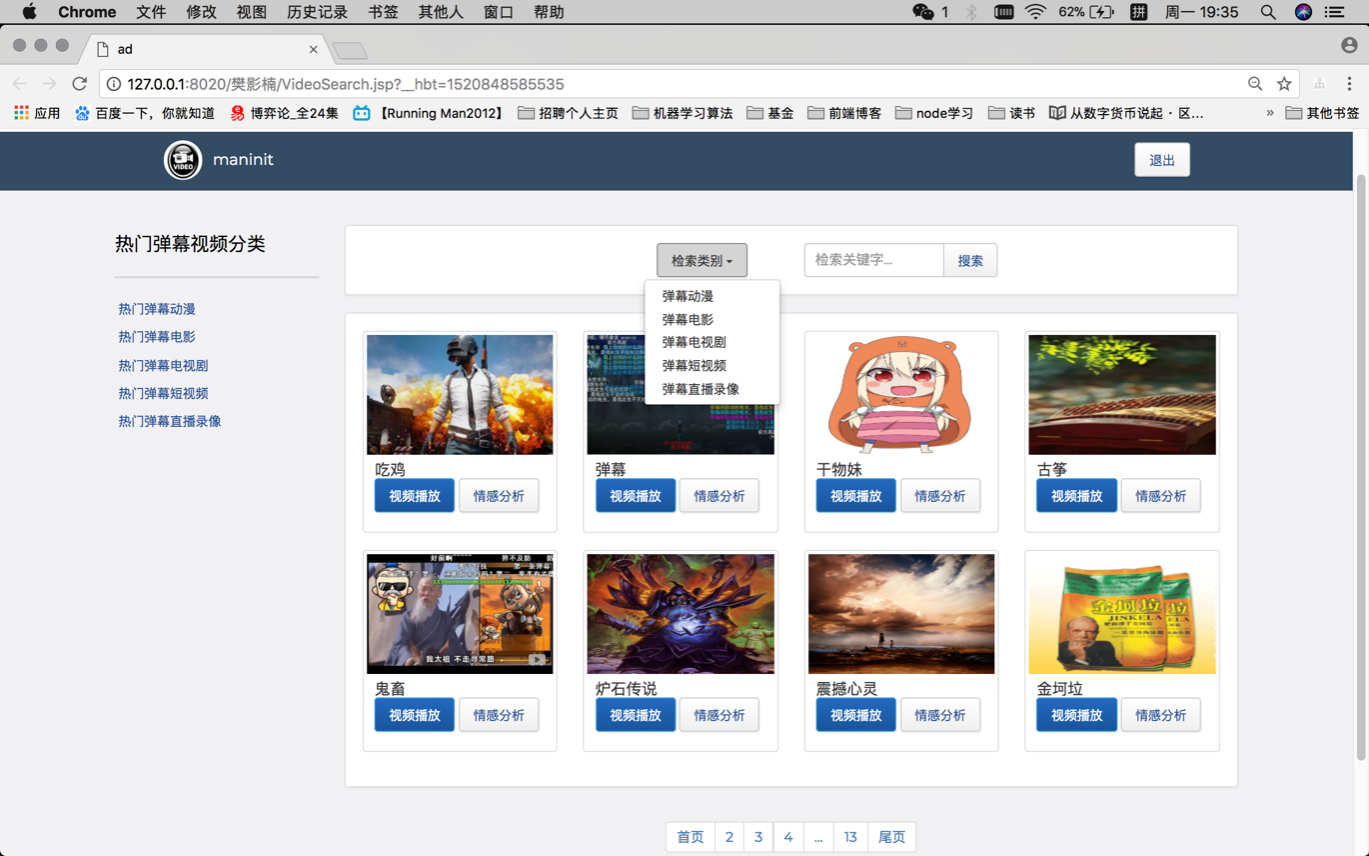


图6-1 视频检索模块系统截图

6.2.2 视频播放模块

1）关键代码

//视频播放

@RequestMapping( value = "videoPlay" )

public ModelAndView videoPlay( @RequestParam("videoID") int videoID, HttpServletRequest request )

{

ModelAndView mv = new ModelAndView( "VideoPlayer" );

Video video = vm.getVideoByVideoID( videoID );

if ( video != null )

{

mv.addObject( "video", video );

}

return(mv);

}

//弹幕播放

//从后端获取弹幕

this.getDanmu = function () {

$.get(that.options.urlToGetDanmu, function (data, status) {

danmuFromSql = eval(data);

for (var i = 0; i < danmuFromSql.length; i++) {

try {

var danmuLs = eval('(' + danmuFromSql[i] + ')');

} catch (e) {

continue;

}

$(that.id + ' .danmu-div').danmu("addDanmu", danmuLs);

}

});

};

if (options.urlToGetDanmu)

this.getDanmu();

//发送弹幕

this.sendDanmu = function (e) {

var text = $(e.data.that.id + " .danmu-input").get(0).value;

if (text.length == 0) {

return;

}

if (text.length > 255){

alert("弹幕过长！");

return;

}

text = text.replace(/&/g, "&gt;").replace(/</g, "&lt;").replace(/>/g, "&gt;").replace(/\"/g, "&quot;").replace(/\n/g, "<br>");

var color = e.data.that.danmuColor;

var position = $(e.data.that.id + " input[name=danmu\_position]:checked").val();

var size = $(e.data.that.id + " input[name=danmu\_size]:checked").val();

var time = $(e.data.that.id + " .danmu-div").data("nowTime") + 3;

var textObj = '{ "text":"' + text + '","color":"' + color + '","size":"' + size + '","position":"' + position + '","time":' + time + '}';

if (e.data.that.options.urlToPostDanmu)

$.post(e.data.that.options.urlToPostDanmu, {

danmu: textObj

});

textObj = '{ "text":"' + text + '","color":"' + color + '","size":"' + size + '","position":"' + position + '","time":' + time + ',"isnew":""}';

var newObj = eval('(' + textObj + ')');

$(e.data.that.id + " .danmu-div").danmu("addDanmu", newObj);

$(e.data.that.id + " .danmu-input").get(0).value = '';

//触发事件

$(e.data.that).trigger("senddanmu");

};

2）实现效果



图6-2 视频播放模块系统截图

6.2.3 弹幕情感分析模块

1）关键代码

弹幕情感分析模块相对代码较复杂，规模较庞大，除了核心的逻辑语句，还使用了很多第三方的库来调用，具体逻辑参照第三章。下面以特征提取和模型训练两个步骤作为例子，将逻辑代码展示如下：

//特征提取模块

var barrageObj = subjectiveBarrageArr[i];

var featureExtractFunc = featureExtractUtil.featureExtract;

var barrageFeatureObj = featureExtractFunc( barrageObj );

/\* 判断：分别将向量提取结果 放入train集和test集中 \*/

barrageObj["sentWordObj"] = barrageFeatureObj.sentWordObj;

barrageObj["degreeWordObj"] = barrageFeatureObj.degreeWordObj;

barrageObj["denyWordObj"] = barrageFeatureObj.denyWordObj;

barrageObj["punctuationObj"] = barrageFeatureObj.punctuationObj;

barrageObj["ngram2"] = barrageFeatureObj.ngram2PosArr;

barrageObj["ngram3"] = barrageFeatureObj.ngram3PosArr;

subjectiveBarrageArr4Train[i]["sentWordObj"] = sentWordObj;

subjectiveBarrageArr4Train[i]["degreeWordObj"] = degreeWordObj;

subjectiveBarrageArr4Train[i]["denyWordObj"] = denyWordObj;

subjectiveBarrageArr4Train[i]["punctuationObj"] = punctuationObj;

/\* 防止没有ngrm2造成ngram属性空 \*/

if ( !ngram2MatchFlag )

{

ngram2PosArr.push( 0 );

ngram2PosArr.push( 0 );

}

subjectiveBarrageArr4Train[i]["ngram2"] = ngram2PosArr;

if ( !ngram3MatchFlag )

{

ngram3PosArr.push( 0 );

ngram3PosArr.push( 0 );

ngram3PosArr.push( 0 );

}

subjectiveBarrageArr4Train[i]["ngram3"] = ngram3PosArr;

//模型训练模块

var classifyDataJSON4Train = fs.readFileSync('./tempData/classifyData4Train.json', 'utf8');

var classifyDataObj4Train = JSON.parse(classifyDataJSON4Train);

var classifyDataJSON4Test = fs.readFileSync('./tempData/classifyData4Test.json', 'utf8');

var classifyDataObj4Test = JSON.parse(classifyDataJSON4Test);

// initialize a new predictor

var clf = new svm.SVM({

svmType: 'C\_SVC',

c: [0.03125, 0.125, 0.5, 2, 8], //c值

// kernels parameters

kernelType: 'RBF',

gamma: [0.03125, 0.125, 0.5, 2, 8],//gamma

// training options

kFold: 4,

normalize: true,

reduce: true,

retainedVariance: 0.99,

eps: 1e-3,

cacheSize: 200,

shrinking : true,

probability : false

});

clf.train(classifyDataObj4Train)

.progress(function(progress){

console.log('training progress: %d%', Math.round(progress\*100));

})

.spread(function (model, report) {

//保存模型到tempData文件中

var modelJSON = JSON.stringify(model);

fs.writeFileSync('./tempData/modelJSON.json', modelJSON);

console.log('training report: %s\nPredictions:', so(report));

var correctNum = 0;

var wrongNum = 0;

var totalNum = 0;

classifyDataObj4Test.forEach(function(ex){

var prediction = clf.predictSync(ex[0]);

if( ex[1] == prediction )

correctNum++;

else

wrongNum++;

totalNum++;

});

console.log("预测总数：%d", totalNum);

console.log("预测正确个数: %d, 预测错误个数: %d", correctNum, wrongNum );

console.log("预测准确率：%f", correctNum/totalNum);

});

2）实现效果



图6-3 弹幕情感分析可视化模块系统截图

6.2.4 视频评论模块

1）关键代码

视频评论模块的主要功能是用户发帖和回复，核心代码如下：

/\*\*

\* 添加留言

\* @param msg

\* @return

\*/

public boolean addMsg(Message msg){

boolean r = false;

String sql = "INSERT INTO message(title,context,wholeft,lefttime)VALUES(?,?,?,?) ";

try{

int num = this.executeUpdate(sql, new String[]{msg.getTitle(),msg.getContext(),msg.getWholeft(),msg.getLefttime()});

if(num > 0){

r = true;

}

}catch(Exception e){

e.printStackTrace();

}finally{

this.closeAll();

}

return r;

}

/\*\*

\* 添加回复

\* @param reply

\* @return

\*/

public boolean addReply(Reply reply){

boolean r = false;

String sql = "INSERT INTO reply(context,messageid)VALUES(?,?)";

try{

int num = this.executeUpdate(sql, new String[]{reply.getContext(),""+reply.getMessageid()});

if(num > 0){

r = true;

}

}catch(Exception e){

e.printStackTrace();

}finally{

this.closeAll();

}

return r;

}

2）实现效果



图6-4 视频评论模块系统截图

6.2.5 个人管理模块

1）关键代码

会员在登陆系统后，可以进行个人管理的相关操作。除了和其他用户之间的私信操作，当用户对系统信息的准确性、系统检索的结果、或系统收录的信息的真实性有所怀疑时，都可以通过私信想管理员进行反馈。用户需要登录自己的账号才能使用该功能模块。用户登录的核心代码如下：

//处理用户登陆校验

@RequestMapping( "checkUserName.htm" )

public void checkUserName( HttpServletRequest request, HttpServletResponse response ) throws Exception

{

String username = request.getParameter( "userName" );

Map < String,Object > map = new HashMap < String, Object > ();

String result = "";

boolean b = loginService.queryByName( username );

if ( b == false )

{

result = "false";

} else {

result = "true";

}

map.put( "result", result );

map.put("dateResult", dateResult);

JSONObject json = JSONObject.fromObject(map);

System.out.println("json "+json);

response.getWriter().print(json.toString());

}

用户登陆账号之后，就可以使用本模块提供的个人管理模块的功能。以下是用个人信息为例的核心代码：

@RequestMapping( "edit\_personalProfile\_action" )

public ModelAndView edit\_personalProfile\_action( HttpServletRequest request )

{

System.out.println( "edit" );

ModelAndView mv = new ModelAndView( "personalProfile" );

User user = um.edit\_personalProfile( request );

mv.addObject( user );

return(mv); /\* 依然跳转至personalProfile页面 \*/

}

2）实现效果

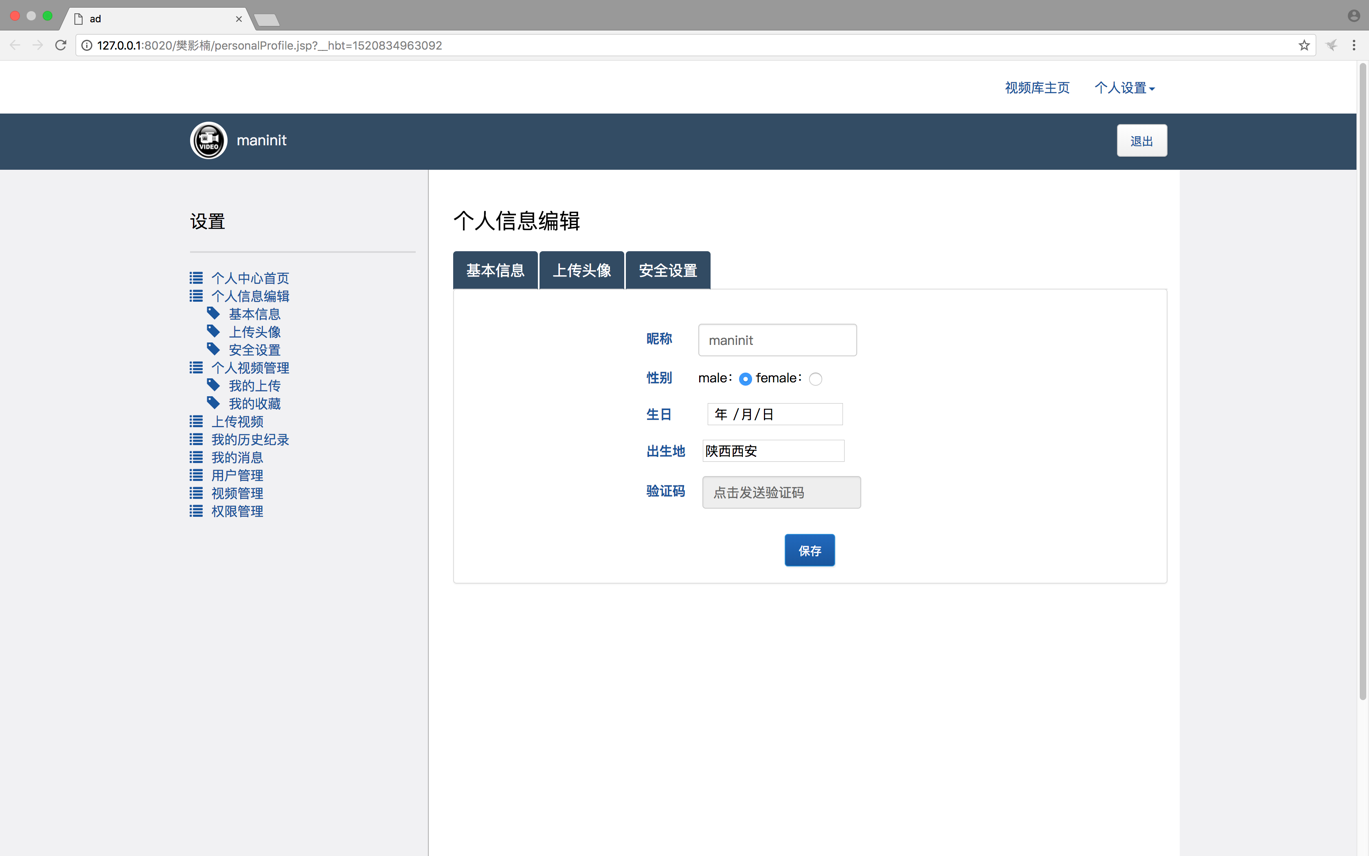


图6-5 个人信息编辑系统截图（一）

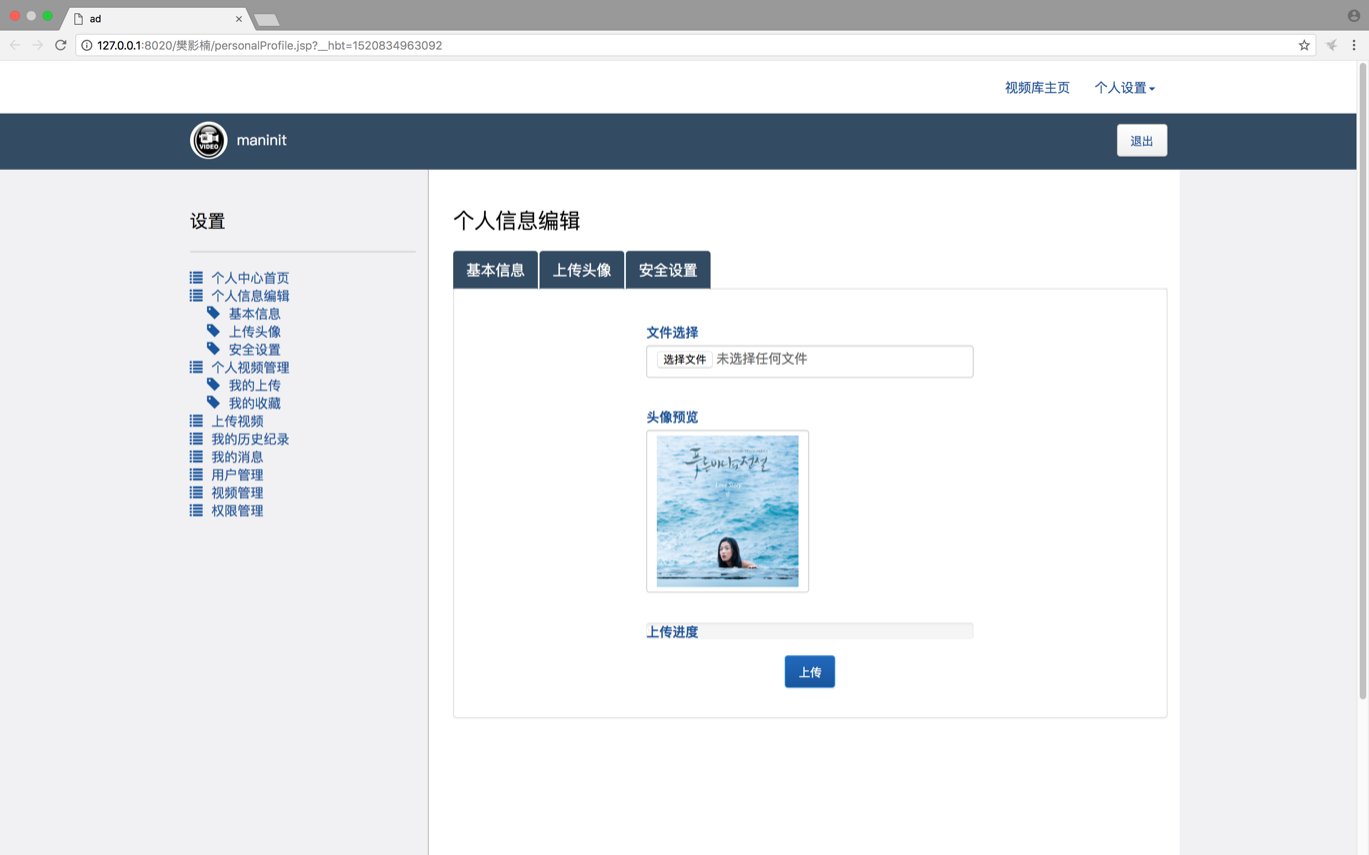


图6-6 个人信息编辑系统截图（二）

## 6.3 系统测试

6.3.1 系统测试环境

软件测试是软件开发过程中的一个重要组成部分，是贯穿整个软件开发生命周期、对软件产品进行验证和确认的活动过程，软件测试的目的是保证本文设计的系统能够正常运行，并且在测试的过程中能够即时发现问题，解决问题。

本文的测试环境如下： 操作系统为Windows7，Java开发环境是jdk1.6.0\_17，以MyEclipse为集成开发环境，以Tomcat6.0作为web服务器，以MySQLServer5.0作为数据库。

6.3.2 系统功能测试

本项目测试完成了对前面需求分析和系统实现描述的各功能模块，对各个功能是否与需求分析一致进行测试，测试用例表和功能测试结果如下表所示：

表6-1 视频查询功能测试用例表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 测试模块 | 测试项目 | 执行步骤 | 测试结果 | 是否通过 |
| 视频查询 | 视频资源浏览 | 进入系统，浏览首页信息 | 系统显示首页资讯 | 是 |
| 视频检索 | 在搜索框中输入关键词，选择检索类型，点击搜索按钮 | 显示结果列表 | 是 |
| 查看视频详情 | 通过首页视频链接或者搜索结果的视频链接，进入视频详情页 | 视频播放和详情浏览 | 是 |

视频播放模块的测试用例表如下表6-2所示：

表6-2 视频播放模块测试用例表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 测试模块 | 测试项目 | 执行步骤 | 测试结果 | 是否通过 |
| 视频及弹幕播放 | 视频自动播放 | 进入视频播放页面 | 视频自动加载和播放 | 是 |
| 视频播放控制（播放/暂停/停止） | 在视频开始正常播放后，点击播放/暂停/停止按钮 | 视频对应的做出相应，即开始播放/暂停播放/停止播放 | 是 |
| 观看弹幕 | 在视频开始正常播放后，弹幕文本在视频浮层上出现 | 弹幕正常出现 | 是 |
| 发送弹幕 | 1.光标选中弹幕输入文本框  2.输入弹幕文本内容  3.点击发送弹幕按钮 | 弹幕发送成功，并显示在视频浮层上 | 是 |

视频情感分析模块测试用例表如下表6-3所示：

表6-3 视频情感分析模块测试用例表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 测试模块 | 测试项目 | 执行步骤 | 测试结果 | 是否通过 |
| 视频评论 | 发帖 | 1.进入视频播放页面  2.鼠标点击视频下方的发帖按钮，弹出发帖框  3.在文本框输入帖子内容  4.点击发布帖子按钮 | 帖子发布成功，并刷新显示在评论区中 | 是 |
| 回复帖子 | 1.选择要回复的帖子，点击回复，弹出回复框  2.在文本框输入帖子内容  3.点击回复帖子按钮 | 回复成功并将回复展示在评论区中 | 是 |
| 收藏帖子 | 选择要回复的帖子，点击收藏按钮 | 帖子被成功收藏 | 是 |
| 删除帖子 | 选择自己待删除的帖子，点击删除 | 帖子被成功删除 | 是 |

表6-3 视频情感分析模块测试用例表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 测试模块 | 测试项目 | 执行步骤 | 测试结果 | 是否通过 |
| 视频情感分析 | 弹幕密度折线图查看 | 进入视频播放页面，等待视频下边弹幕密度折线图加载 | 弹幕密度折线图加载并显示 | 是 |
| 弹幕情感值波动折线图查看 | 进入视频播放页面，等待视频下边弹幕情感波动折线图加载 | 弹幕情感波动折线图加载并显示 | 是 |
| 弹幕情感折线图点细节数据查看 | 弹幕情感折线图加载成功之后，鼠标移动到折线图的节点上 | 鼠标旁弹出数据详情内容框，展示该点的情感数据 | 是 |

视频评论模块测试用例表如下表6-4所示：

表6-4 视频评论模块测试用例表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 测试模块 | 测试项目 | 执行步骤 | 测试结果 | 是否通过 |
| 视频评论 | 评论查看 | 进入视频播放页面，等待视频下边评论信息加载 | 评论信息加载并显示 | 是 |
| 发布评论 | 进入视频播放页面，在评论输入框输入评论并点击发送 | 评论发布成功并刷新评论区 | 是 |
| 评论回复 | 选中评论对象并点击回复，在弹出的回复框中输入回复，并点击回复按钮 | 回复成功并刷新评论区 | 是 |

个人管理模块测试用例表如下表6-5所示：

表6-5 个人管理模块测试用例表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 测试模块 | 测试项目 | 执行步骤 | 测试结果 | 是否通过 |
| 个人管理 | 系统初始化 | 点击页面上方的个人信息 | 显示系统首页 | 是 |
| 用户注册 | 1.点击页面右上角的注册，显示注册页面  2.在注册页面填写用户名、登陆密码等必填信息，点击确定  3.弹出注册成功提示窗口，点击确定 | 显示登录首页 | 是 |
| 正常登录 | 1.点击页面右上角的登录按钮，显示登录页面  2. 在登录页面填写用户名、登录密码信息和选择用户类型，点击确定。 | 按照用户的类型显示系统登录后的页面 | 是 |
| 错误登录 | 1.点击页面右上角的登录图标，显示登录页面； 2.在登录页面填写错误的用户名、登录密码信息，或者填写与用户名不匹配的用户类型，点击确定。 | 但除提醒用户名或密码错误的提示窗口 | 是 |
| 修改个人信息 | 1.点击页面上方的个人信息； 2.在显示的个人信息页面选择修改按钮，填写需要修改的内容，点击确定。 | 个人信息页面显示修改以后的信息 | 是 |
| 查看个人信息 | 点击页面上方的个人信息 | 显示个人信息页面 | 是 |
| 删除收藏帖子 | 在会员登录条件下点击页面上方的收藏夹管理  在帖子管理页面点击查看帖子信息  选择要删除的帖子，在其后面的删除按钮上点击确定 | 删除所选帖子，显示剩余帖子的信息 | 是 |

6.3.3 系统性能测试

性能测试主要对响应时间、事务处理速率和其他与时间相关的需求进行评测和评估。性能评测的目标是核实在正常的预期工作量和预期的最繁重工作量情况下性能需求是否都已满足。由需求分析得到的性能需求指标如下： 并发用户数小于1000，在1000个并发用户进行股票检索时，业务处理响应时间在5秒以内；对股票进行交易时，不计入网络传输时间，买进和卖出的响应时间在3秒以内。当使用股票论坛功能时，回帖、发帖的响应时间应该在5秒以内。 系统实现后，在搭建的LoadRunner测试环境中，让系统的服务器一直处于开启状态，让系统持续运行了3天，没有发生意外中断的情况，可以有效工作，客观地证明了系统满足稳定性指标。性能测试的结果如下表6-6所示（秒：s）。

表6-6 系统性能测试表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 并发用户数 | 事务平均响应时间 | | | 事务最大响应时间 | | |
| 视频检索 | 视频及弹幕播放 | 视频评论 | 视频检索 | 视频及弹幕播放 | 视频评论 |
| 50 | 1.0s | 0.6s | 0.9s | 1.4s | 0.8s | 1.1s |
| 100 | 1.3s | 0.7s | 1.0s | 1.6s | 1.0s | 1.2s |
| 200 | 1.6s | 0.9s | 1.3s | 1.8s | 1.1s | 1.6s |
| 500 | 2.2s | 1.2s | 1.5s | 2.6s | 1.5s | 1.9s |
| 1000 | 3.7s | 1.8s | 2.6s | 4.3s | 2.4s | 4.1s |

在系统性能测试的的过程中，发现100个以内用户同时使用系统时，系统的处理速度变化不大；200个以内用户同时使用系统时，系统能够快速准确地完成业务的处理；在200至500个用户同时使用系统时，系统的处理时间加长，并且有处理失败的情况，但系统仍能正常运行。当并发数达到1000时，发帖的错误率有一定上升，并出现检索失败的情况，但系统人能够正常运行。测试说明该系统达到了支持1000个用户同时操作的目标，保证了系统在需求范围内正常工作，满足了业务处理能力指标的要求。

## 6.4 本章小结

本章将基于情感分析的视频弹幕系统的实现的核心代码、功能截图以及测试过程进行详细的展示，以描述系统的实现与测试。在代码与功能展示上以模块为单位，具体描述了系统视频检索模块、视频播放模块、弹幕情感分析可视化展示模块、视频评论模块、个人管理模块、管理员管理模块的核心代码和系统界面。随后，本章对系统的功能进行了用例测试，保证了系统的实用性。

# 7 结论与展望

### 7.1 总结

在线视频这种web信息媒介已经成了我们日常获取资讯的一种重要途径。如何更好的通过自动化的方式对视频的内容进行解析，进而给视频观看用户更好的观看体验，这成了目前很多学者和领域工作者研究的着眼点。弹幕文本作为近几年新兴的一种评论方式，有着和视频时间点高度相关的特性，是一种用来从侧面解析视频的内容以及剧情波动的一种很有效的手段，会在当今一段时间内带来研究热潮。本文在这一背景下，针对弹幕短文本的特点，提出了一种改进的文本情感分析方法，并设计实现了一个基于情感分析的视频弹幕系统。本文的具体工作如下：

1）针对弹幕文本长度较短，特征较稀疏的特点，本文提出了一种结合高维混合特征的SVM情感分析方法。模型着重在文本特征提取环节，将文本语义特征，词性特征等有效特征进行比对和整合，提出了一种新的情感特征提取方法。与此同时，本文针对弹幕文本的客观语料含有大量情感信息的特点，提出了一种基于客观语料的情感分类方法。通过识别弹幕密集区，并将主客观语料分类，将主观语料的情感计算值映射给客观语料区，从而通过反复训练情感分类模型，提高弹幕文本情感分类的准确度。

2）在对视频弹幕文本的情感分析之后，本文涉及并实现了一个基于弹幕文本的视频弹幕系统。该系统可以为用户提供热门视频资源的检索，并进行视频和弹幕的播放，用户可以发布自己的弹幕进行互动。此外，用户还可以再系统的评论中进行交流，发布自己对于视频的想法和感受。本文从系统的徐局分析、建模设计、实现测试三个角度对系统的构建过程进行详细的描述，并对系统按照测试用例进行了测试、给出运行结构。测试结果表明，本系统能够满足用户对视频弹幕系统的需求。

### 7.2 展望

本文针对视频弹幕的特点，提出了一种针对弹幕文本的情感分析算法，并设计、实现了一个基于视频弹幕的视频平台系统。然而，由于自身的知识和能力的限制，本文在算法设计和系统实现上仍然存在一些不足，具体如下：

1）在对弹幕的分析上，由于之前没有开源的弹幕文本标准数据集，本文选取的数据来源是自己基于人工标注的弹幕情感语料，数量有限。而事实上在计算性能高度发展的今天，这样规模的标注数据作为研究的数据基础已经远远不够。希望以后在数据集上能够更加丰富，也只有这样才能够为用户提供能加有价值的信息。

2）弹幕文本情感分析算法中，我们主要选用了基于词典的方法和基于支持向量机的情感分类方法作为原型进行改进，缺乏和其他方法的对比，算法实验中也缺少多种算法的分析结果对比，从而导致本文提出的算法模型可能并不是问题的最优解，在其他分类模型下可能有更大的改进空间。

3）在视频弹幕平台这个系统中，本系统只能提供视频弹幕的倾向性分析，对视频的分析角度过于单一，没有结合视频本身的内容去做分析，从而不能从更高的角度为用户的视频观看提供决策帮助。由于时间和专业技能有限，因此在有限的时间内无法进行深入的研究。

4）从系统的设计与实现看，系统在测试阶段没有进行压力测试，因此无法确定在大流量访问下的性能是否稳定。

针对以上的问题，作者将在后续的工作中从算法和系统两个角度做出改进。算法上，将会考虑更多文本分类模型，对多种分类方法进行对比验证。同时，将对系统进行更多测试，以便于及时发现问题，完善本文的工作。

# 致谢

三年的研究生生活很快就要结束了，在我读研和撰写论文的过程中，得到了很多人的帮助，在此表示衷心的感谢，并致以我深深的敬意。

我最需要感谢的是我的论文指导老师——饶元老师。从科研课题的确立到指导科研阶段，再到确定初期的整个论文的思路和方向，直至论文的精心审阅，饶老师都认真负责地指导我、教育我，不仅使我在技术上有了明显的进步，而且思想上提升了高度。尤其是在我撰写论文的过程中，饶老师对论文严格要求，多次指导，监督论文进度并提出很多有用的意见，使我受益匪浅。饶老师治学态度严谨，工作一丝不苟，对学术精益求精，这都将成为我以后工作和学习的榜样。感谢饶老师对我的悉心指导和热情鼓励，向饶老师致以我深深的谢意。

我还要感谢软件学院的各位领导和老师，是他们给予了我很多指导和帮助，为我们创造了和谐的学习氛围。谢谢各位老师们。

感谢我的父母及家人，感谢他们为我默默地奉献，支持我，关心我，在我遇到困难的时候给我信心和力量，这是我一生都难以回报的。感谢我的同学和朋友们，他们对我无微不至的关心和支持，给了我前进的动力。

感谢给予参考和引用权的资料、图片、文献、研究成就的作者，是他们的作品指导我在这一领域有了更大的研究和进步。

最后感谢参加论文评审和答辩的各位老师们，谢谢老师对本文的批评指正。

# 参考文献

1. 王义真,郑啸,后盾,胡昊. 基于SVM的高维混合特征短文本情感分类[J]. 计算机技术与发展,2018,(02):1-6.
2. 李杰,李欢. 基于深度学习的短文本评论产品特征提取及情感分类研究[J]. 情报理论与实践,,:1-9.
3. Alexander Pak, Patrick Paroubek. Microblogging for Micro Sentiment Analysis and Opinion Mining Le microblogging pour la micro analyse des sentiments et des opinons[J]. Traitement Automatique des Langues,2011,51(3):.
4. 沈冀,马志强,李图雅,张力. 面向短文本情感分析的词扩充LDA模型[J]. 山东大学学报(工学版),,:1-7.
5. 程传鹏,苏安婕. 一种短文本特征词提取的方法[J]. 计算机应用与软件,2014,(06):162-164+212.
6. Wingyan Chung,Daniel Zeng. Social‐media‐based public policy informatics: Sentiment and network analyses of U . S . Immigration and border security[J]. Journal of the Association for Information Science and Technology,2016,67(7):.
7. Pintu Lohar,Haithem Afli,Andy Way. Maintaining Sentiment Polarity in Translation of User-Generated Content[J]. The Prague Bulletin of Mathematical Linguistics,2017,108(1):.
8. 刘勘,袁蕴英. 基于自动编码器的短文本特征提取及聚类研究[J]. 北京大学学报(自然科学版),2015,(02):282-288.
9. 全湘溶. 中文短文本多级情感分析[J]. 现代电信科技,2015,(05):51-59.
10. 郭永辉. 面向短文本分类的特征扩展方法[D].哈尔滨工业大学,2013.
11. Pilar Corredor,Elena Ferrer,Rafael Santamaria. Sentiment-prone investors and volatility dynamics between spot and futures markets[J]. International Review of Economics and Finance,2015,35:.
12. 黄文明,孙艳秋. 基于最大熵的中文短文本情感分析[J]. 计算机工程与设计,2017,(01):138-143.
13. 邓扬,张晨曦,李江峰. 基于弹幕情感分析的视频片段推荐模型[J]. 计算机应用,2017,(04):1065-1070+1134.
14. 唐晓波,刘广超. 细粒度情感分析研究综述[J]. 图书情报工作,2017,(05):132-140.
15. Steve Y. Yang,Sheung Yin Kevin Mo,Anqi Liu. Twitter financial community sentiment and its predictive relationship to stock market movement[J]. Quantitative Finance,2015,15(10):.
16. Pratik Thakor,Sreela Sasi. Ontology-based Sentiment Analysis Process for Social Media Content[J]. Procedia Computer Science,2015,53:.
17. 张婧婧,杨业宏,安欣. 弹幕视频中的学习交互分析[J]. 中国远程教育,2017,(11):22-30+79-80.
18. 李科. 基于多元特征融合和LSTM神经网络的中文评论情感分析[D].太原理工大学,2017.
19. 王骏科. 多粒度中文情感分析系统的研究与实现[D].电子科技大学,2017.
20. Felix Greaves,Daniel Ramirez-Cano,Christopher Millett,Ara Darzi,Liam Donaldson. Machine learning and sentiment analysis of unstructured free-text information about patient experience online[J]. The Lancet,2012,380:.
21. 罗好. 视频网站弹幕视频传播影响因素研究[D].哈尔滨工业大学,2017.
22. 郑飏飏,徐健,肖卓. 情感分析及可视化方法在网络视频弹幕数据分析中的应用[J]. 现代图书情报技术,2015,(11):82-90.
23. 宋静静. 中文短文本情感倾向性分析研究[D].重庆理工大学,2013.
24. Maribel Yasmina Santos,Carina Sofia Andrade. Sentiment Analysis with Text Mining in Contexts of Big Data[J]. International Journal of Technology and Human Interaction (IJTHI),2017,13(3):.
25. 杜振雷. 面向微博短文本的情感分析研究[D].北京信息科技大学,2013.
26. 刘楠. 面向微博短文本的情感分析研究[D].武汉大学,2013.
27. 袁丁. 中文短文本的情感分析[D].北京邮电大学,2015.
28. 沈磊. 基于规则与机器学习方法的中文微博情感分析研究[D].安徽大学,2015.
29. 陈文. 中文短文本跨领域情感分类算法研究[D].重庆大学,2016.
30. Mazen El-Masri,Nabeela Altrabsheh,Hanady Mansour,Allan Ramsay. A web-based tool for Arabic sentiment analysis[J]. Procedia Computer Science,2017,117:.
31. 张慧. 用户生成内容情感分析方法研究[D].安徽财经大学,2016.
32. 刘续乐,何炎祥. 基于多特征的微博情感分析研究[J]. 计算机工程,2017,(12):160-164+172.
33. Wang,Zheng. Sentiment classification of Chinese online reviews: a comparison of factors influencing performances[J]. Enterprise Information Systems,2016,10(2):.
34. 张璞. Web评论文本情感分类方法研究[D].重庆大学,2015.
35. 朱晓光. 基于半监督学习的微博情感分析方法研究[D].山东财经大学,2014.
36. 何慧. WEB文本挖掘中关键问题的研究[D].北京邮电大学,2009.
37. 何天翔. 基于情感词网的短文本情感分类方法研究[D].西南科技大学,2015.
38. 葛文镇. 面向微博的短文本多分类研究[D].宁波大学,2015.
39. 张英. 基于深神经网络的微博短文本情感分析研究[D].中原工学院,2017.
40. 李洋. 面向中文产品评论数据的情感分析模型设计及评估[D].北京邮电大学,2017.
41. 叶强,张紫琼,罗振雄. 面向互联网评论情感分析的中文主观性自动判别方法研究[J]. 信息系统学报,2007,(01):79-91.
42. 张紫琼. 面向中文情感分析的词类组合模式研究[D].哈尔滨工业大学,2007.
43. 金鑫. 基于朴素贝叶斯的文档级情感分析[D].大连理工大学,2013.

# 攻读学位期间取得的研究成果

学位论文独创性声明（1）

本人声明：所呈交的学位论文系在导师指导下本人独立完成的研究成果。文中依法引用他人的成果，均已做出明确标注或得到许可。论文内容未包含法律意义上已属于他人的任何形式的研究成果，也不包含本人已用于其他学位申请的论文或成果。

本人如违反上述声明，愿意承担以下责任和后果：

1．交回学校授予的学位证书；

2．学校可在相关媒体上对作者本人的行为进行通报；

3．本人按照学校规定的方式，对因不当取得学位给学校造成的名誉损害，进行公开道歉。

4．本人负责因论文成果不实产生的法律纠纷。

论文作者（签名）： 日期： 年 月 日

学位论文独创性声明（2）

本人声明：研究生 所提交的本篇学位论文已经本人审阅，确系在本人指导下由该生独立完成的研究成果。

本人如违反上述声明，愿意承担以下责任和后果：

1．学校可在相关媒体上对本人的失察行为进行通报；

2．本人按照学校规定的方式，对因失察给学校造成的名誉损害，进行公开道歉。

3．本人接受学校按照有关规定做出的任何处理。

指导教师（签名）： 日期： 年 月 日

学位论文知识产权权属声明

我们声明，我们提交的学位论文及相关的职务作品，知识产权归属学校。学校享有以任何方式发表、复制、公开阅览、借阅以及申请专利等权利。学位论文作者离校后，或学位论文导师因故离校后，发表或使用学位论文或与该论文直接相关的学术论文或成果时，署名单位仍然为西安交通大学。

论文作者（签名）： 日期： 年 月 日

指导教师（签名）： 日期： 年 月 日

(本声明的版权归西安交通大学所有，未经许可，任何单位及任何个人不得擅自使用)