摘 要

本项目意在实现一款具有实际应用价值的影视票房预测系统。在实现的过程中，为了提高预测模型的准确率，借鉴了影视作品的基本信息，并基于社交媒体上的影视评论，开展了影评的情感分析和意见挖掘工作，从而抽取影片口碑信息，并创造性地提出影人主创指数的概念。最终，综合以上指数和数据维度，有效地实现了影视票房的分类预测。

在口碑分析实现的过程中，提出了综合表情符号、核心情感词典、分类器的情感判定方法，能够有效实现影评积极、消极两级情感判定和积极、中性、消极三级情感判定模型。其中，两级情感分析达到82.2%的准确率，三级情感分析达到67.5%的准确率，均达到领域较高水平。同时，辅助基于TextRank算法的影评关键词分析，挖掘影评中的意见信息，最终得出影片的口碑情况。

在分析影片主创团队时，提出了影人主创指数的概念，统计影人在近年来以不同身份参与影片取得的票房情况，结合时间权重和角色权重，计算影人主创指数，分析影人的票房影响力，从而有效地将影人信息融入到影视票房预测中来。

最终，影视票房预测，基于2011~2016年的影视作品信息，综合影片的基本信息、主创团队主创指数、影片口碑等多个数据维度，借助机器学习的方式，提出了影视票房预测模型，能够实现对影片总票房百万以下、百万、千万、亿、十亿的分类预测，并达到70%准确率，具备行业实际应用价值。

关键词： 影评情感分析；主创指数；票房预测；机器学习；TextRank

Abstract

The project is intended to achieve a film box office prediction system with practical value. In the process of implementation, in order to improve the accuracy of prediction models, in addition to referring to film works basic information, but also based on video reviews, carrying out the sentiment analysis and opinion mining work, and effectively obtain video reputation information. Besides, put forward the concept of filmmakers’ creative index originally. Summing up these factors, realize a film box office prediction system finally with high accuracy.

In the process of realization of video reputation analysis, this paper put forward a comprehensive method based on emoticons, core emotional dictionary, and emotion classifiers. The method can effectively achieve positive, negative emotional judgment with 82.2% accuracy rate, and positive, neutral, negative emotion judgement which can reach 67.5% accuracy rate. At the same time, the auxiliary critic keyword analysis based TextRank algorithm, can mining critic of opinion information, and obtain word of mouth of movie finally.

When analyzing the video creative team, the concept of filmmakers’ creative index is proposed, it needs to statistics the box office which filmmakers made over the past few years at different times in different role to participate in film. Combined with time weight and role weight, calculate filmmakers creative index, analyze filmmakers’ influence on the box office, so that the filmmakers information can better integrated into the film box office prediction.

Ultimately, the movie box office prediction carried out by the data set of 2011 to 2016 films. Considering basic information about the integrated video, creative team creative index, film reputation info, etc., and using machine learning algorithm, the film box office prediction model can effectively predict box office class and reach 70% accuracy rate.

**Keywords**: Sentiment analysis of film reviews, Filmmakers’ creative index, Box office Predict, Machine Learning, TextRank

目 录

[摘 要 I](#_Toc455066802)

[Abstract II](#_Toc455066803)

[第1章 绪论 1](#_Toc455066804)

[1.1 项目来源 1](#_Toc455066805)

[1.2 项目开发目的和意义 1](#_Toc455066806)

[1.3 国内外相关领域现状分析 2](#_Toc455066807)

[1.4 本文主要工作内容 3](#_Toc455066808)

[第2章 影视票房预测系统需求分析及总体设计 5](#_Toc455066809)

[2.1 影视票房预测系统需求分析 5](#_Toc455066810)

[2.1.1 影视票房预测系统业务需求 5](#_Toc455066811)

[2.1.2 影视票房预测系统功能需求 5](#_Toc455066812)

[2.1.3 影视票房预测系统非功能需求 8](#_Toc455066813)

[2.2 影视票房预测系统总体设计方案 9](#_Toc455066814)

[2.2.1 影视票房预测系统架构设计 9](#_Toc455066815)

[2.2.2 影视票房预测系统功能结构设计 10](#_Toc455066816)

[2.3 本章小结 11](#_Toc455066817)

[第3章 影视票房预测系统详细设计 12](#_Toc455066818)

[3.1 影视票房预测系统存储设计 12](#_Toc455066819)

[3.1.1 社交影评及用户信息存储设计 12](#_Toc455066820)

[3.1.2 影视作品信息存储设计 14](#_Toc455066821)

[3.2 影视票房预测系统核心算法设计 17](#_Toc455066822)

[3.2.1 影评情感分析算法设计 17](#_Toc455066823)

[3.2.2 影评关键词提取算法设计 20](#_Toc455066824)

[3.2.3 主创指数算法设计 21](#_Toc455066825)

[3.2.4 受众人群分析算法设计 23](#_Toc455066826)

[3.2.5 影视票房预测算法设计 24](#_Toc455066827)

[3.3 影视票房预测系统功能实现详细设计 26](#_Toc455066828)

[3.3.1 影评情感分析实现详细设计 26](#_Toc455066829)

[3.3.2 影人主创指数实现详细设计 28](#_Toc455066830)

[3.3.3 影视票房预测实现详细设计 29](#_Toc455066831)

[3.4 本章小结 32](#_Toc455066832)

[第4章 影视票房预测系统实现 33](#_Toc455066833)

[4.1 影视评论分析实现 33](#_Toc455066834)

[4.1.1 影评情感分析实现 33](#_Toc455066835)

[4.1.2 影评关键词提取实现 35](#_Toc455066836)

[4.2 影视作品分析实现 37](#_Toc455066837)

[4.2.1 影人主创指数实现 37](#_Toc455066838)

[4.2.2 影视作品受众人群分析实现 38](#_Toc455066839)

[4.2.3 影视作品票房预测实现 39](#_Toc455066840)

[4.3 系统整体可视化呈现 42](#_Toc455066841)

[4.4 本章小结 45](#_Toc455066842)

[第5章 影视票房预测系统测试与性能分析 46](#_Toc455066843)

[5.1 影视票房预测系统功能测试 46](#_Toc455066844)

[5.1.1 影评情感分析功能测试 46](#_Toc455066845)

[5.1.2 影评关键词提取功能测试 46](#_Toc455066846)

[5.1.3 影人主创指数功能测试 47](#_Toc455066847)

[5.1.4 影片票房预测功能测试 48](#_Toc455066848)

[5.2 影视票房预测系统核心模块性能分析 48](#_Toc455066849)

[5.2.1 影评情感分析性能分析 48](#_Toc455066850)

[5.2.2 影视票房预测性能分析 49](#_Toc455066851)

[5.3 本章小结 51](#_Toc455066852)

[结 论 52](#_Toc455066853)

[参考文献 53](#_Toc455066854)

[哈尔滨工业大学本科毕业设计（论文）原创性声明 55](#_Toc455066855)

[致 谢 56](#_Toc455066856)

[附录1 外文文献译文 57](#_Toc455066857)

[附录2 外文文献原文 69](#_Toc455066858)

# 绪论

本章主要介绍项目来源以及项目开发的目的与意义，并简要分析了国内外在相关领域上的研究应用现状。最后，综合全文的主要内容，进行概括性综述，以便快速地了解本项目的概要情况。

## 项目来源

本项目来源于微瑞思创有限公司影视价值分析的内部产品需求。微瑞思创是一家以大数据技术为基础的信息技术公司，其现有产品“影视宝”，是一款集影视作品数据收集、呈现、分析等多种功能的影视垂直行业产品。为了进一步提升“影视宝”产品的核心竞争力，以及满足优酷、DMG等客户需求，因此开展影视数据分析等相关工作，通过数据挖掘的方式为行业提供更好的影视价值分析解决方案。

## 项目开发目的和意义

影视作品分析及票房预测一直以来都是影视行业备受关注的问题，其对于影视作品的投资和收益具有重大的指导意义。从影视初期IP产权的选择投资，到影视上映后，优酷、爱奇艺等在线媒体对于影视作品二次投资和购买等，都离不开对影视价值评估这一重要一环。而传统的影视价值评估，多是通过影视行业投资人员及业内专家凭借行业经验开展的，如何从琳琅满目的热门IP中选择投资的对象，如何在影视作品宣传期间有针对性的投放宣传亮点，如何设计作品的先行预告片，以及如何投资和发展影视作品的周边经济等，就都成为了这个行业至关重要的人为决策问题。因此，如何通过信息化的技术，基于影视数据的客观信息及社交媒体的大众评论数据，提出一套完整的影视作品分析模型，为行业人员的决策提供指标和理论依据，成为了一件非常开创意义和商业价值的事情。

此外，公司目前主要为企业提供大数据产品、技术、营销及分析咨询服务。尤其是在影视行业，其现有产品“影视宝”，已经日趋成型，集成了影视作品信息收集、分析、可视化呈现等多种功能，并在影视行业俘获了大批的客户群体，建立了良好的口碑。因此开展影视分析及票房预测等相关工作，丰富和完善“影视宝”产品的功能，对于公司产品核心竞争力的提升具有重要的战略意义，进而俘获更多的用户，巩固产品在行业领域的领先地位。

## 国内外相关领域现状分析

近几年来，随着社交媒体不断渗透到人们的日常生活，越来越多的人通过社交平台表达对现实世界的意见和见解。基于社交媒体的评论，可以分析商品的使用价值、社会舆论动向、用户情感倾向等[1]。

在近期的研究中，也有不少研究开展了基于社交媒体上的用户活动预测影视票房的相关工作。Jingfei Du等人在预测影视票房的过程中，提出了统计影视评论的转发量和评论量，作为影视票房分析的数据特征的想法，同时结合评论的情感分类和历史票房，最终达到预测票房的目的[2]。史伟则提出了结合情感词特征的影视票房自回归模型，也取得了较好的效果[3]。而在影视价值的影响因素分析上，胡小莉等人则通过模型验证和相关度分析，得出了导演、演员、续集、翻拍、地区、档期、类型等数据特征对于影视价值分析具有较大影响[4]。因此，能够综合影视作品各方面的信息，实现影视作品价值评估和票房预测具备一定的实验难度和应用价值。

在具体工作开展过程中，除了影视价值预测，系统另外一个难点则在于微博评论的情感分析模型，透过大众评论，挖掘用户的喜好和好评情况，从而为影视价值评估提供更有利的依据。关于影视评论的情感分析，国内外则开展过很多相关工作。常见的语义情感倾向分析主要从两个角度入手，一个是基于情感词和句子语法结构的角度分析语句的情感倾向，另一个角度便是收集大量的训练语料，抽取特定的数据特征，结合机器学习的方法，开展学习和训练。

基于情感词的分析，英文的相关工作可以参考Mike Thelwall等人的实验，主要采用了基于情感核心词典的方式，对于词典中的每一个词，不仅做情感倾向的判定，同时对情感倾向度进行5级打分，分析过程中考虑修饰副词及否定词等的影响，基于分句并行地对句子正逆向情感进行打分，最终达到情感分析的目的，其准确率可达到72.8%[5]。类似的中文分析，则可以参考Jingang Wang等同学的工作，针对微博评论，基于基础核心词典，开展SO-PMI的扩词，分析过程中，借助语法树的构建，定位句子的词根，找出周边词对核心词的修饰关系，最终通过核心词为句子的情感倾向进行打分，最终可达到63.6%的准确率[6]。

而基于机器学习的方式，则在自然语言处理领域，主要采用的机器学习方法有支持向量机、朴素贝叶斯、最大熵等模型，常用的特征除了常见的unigram, bigram, trigram等，还会引入情感值、语句标点、词性信息、语法树结构等辅助信息，进而丰富机器学习的特征，提高情感预测的准确率。这方面工作可以参考早期的Bo Pang等人的工作，在其分析过程中，对于SVM、Naïve Bayes、Max Ent等常用算法均作了尝试，采用了基本的unigrams结合bigrams的方式，最终两级情感划分达到了82.7%的准确率[7]。类似地，中文方面也可以参考刘鲁、刘志明博士的工作，其在特征选择上考虑的TF-IDF以及IG的方式，结合SVM和朴素贝叶斯，也得到了良好的效果，准确率可以达到85%以上。后续研究中，针对英文评论，结合SVM算法，bigram特征以及情感词典特征，最终基本可以达到90%的准确率[8]。而在后期的情感分析过程中，两级情感划分已经无法满足日常的应用需求，因此更多的多级情感划分以及多类情感分析等相关工作也由此展开，多级情感如Theresa Wilson等人开展的情感强度挖掘的相关工作，结合语法依赖树，构建数据特征，开展情感意见的四级判定，最终也达到了近似57%的准确率[9]。多类情感分析，则可以参考董喜双等老师的工作，对准喜、怒、哀、惧四类情感分析问题，重点解决了中文词、句的情感分析问题[10]。将词的情感分析处理为候选词情感分类问题。首先通过词性过滤获得候选词，进而根据特征模板获取候选词情感特征，然后应用最大熵模型判断候选词情感类别，并通过情感词结合最大熵模型对语句的情感也做了判定，也得到了较好的效果。

## 本文主要工作内容

本文主要研究的内容是影视作品分析及票房预测，收集海量影视作品信息，分析相应的受众人群，口碑情况，以及最终提出具有可行性的影视票房预测方案并投入实际使用，辅助行业决策。因此在具体开展过程中，主要开展的工作有：

（1）爬取微博影评，开展数据标注和模型训练，最终提出完善的三级情感判定模型，分析影视作品的大众情感倾向及口碑情况；

（2）收集发表微博影评的相应用户信息，分析影视作品的受众分布；

（3）针对影视作品的主创成员，开展票房影响力分析，最终提出主创指数，辅助影视票房预测；

（4）综合影视作品基本信息、主创指数、口碑信息，开展模型训练，提出影视票房预测模型；

为了描述上述工作内容，本文将分为如下5个章节：

第1章 绪论 绪论章节主要介绍了项目的概要情况，帮助阅读者快速了解项目背景，项目意义以及国内外相关领域的开发和应用现状；

第2章 需求分析及总体设计方案 第2章对系统进行了需求分析和总体设计，主要确认了系统的业务需求以及功能需求，并在此基础上对系统进行了功能结构设计和模块初步设计。

第3章 影视票房预测系统详细设计 第3章对系统进行了详细设计，设计了系统的数据结构以及主要算法，确认了系统的主要流程。

第4章 影视票房预测系统实现 第4章则主要对系统核心功能点影评情感分析、关键词提取、影视票房预测、受众人群分析的详细实现流程进行了说明，并详细展示了相应的实现效果。

第5章 影视票房预测系统测试与性能分析 第5章对系统进行了测试和性能分析，并给出了主要的测试用例以及测试输出结果，并针对系统的核心功能模块和算法，进行了性能对比和分析。

# 影视票房预测系统需求分析及总体设计

本章首先对本项目进行了需求分析，基于本系统的业务需求，深入地描述了系统的功能需求和非功能性需求。在此基础上，对整个系统的架构、模块划分、功能结构进行了整体的设计，从而为后续的系统开发打下坚实的基础，提高开发效率并确保其正确性。

## 影视票房预测系统需求分析

### 影视票房预测系统业务需求

影视作品数据分析，主要包括影视作品口碑分析、受众人群分析、主创团队影响力分析、票房预测、可视化呈现等业务需求，具体业务需求描述如下：

（1）口碑分析 基于微博等社交媒体上的评论，进行情感倾向分析和意见挖掘，获取影片的大众反馈，分析大众对于影视作品的情感倾向；

（2）受众人群分析 分析影片的受众群体，从年龄、收入、学历、性别等多个角度出发，分析影片受众年龄分布、收入分布、学历分布、性别分布、兴趣爱好分布等，找到影片真正的目标人群，为后续影视作品的投资和宣传提供有力的支撑；

（3）主创团队影响力分析 分析影视作品主创团队的影响力，从主创成员历史作品的票房纪录出发，综合考虑作品发行的时间和主创成员参与的角色身份，分析影人的票房影响力，最终得出整个主创团队的票房号召力；

（4）影视票房预测 影视作品票房预测，则需要考虑众多因素和数据维度，需要从影视作品基本信息出发，结合主创团队的影响力、预告片的播放情况、社交平台口碑指数等，最终实现影视票房预测功能，对未上映的影片票房实现有效的分档预测。

（5）可视化呈现 将影片的基本信息、票房排片情况、口碑信息、预测票房、受众分布、主创社交账号粉丝属性等进行可视化呈现，完整的展现影视作品各个方面的信息。

### 影视票房预测系统功能需求

为了实现上面提出的业务需求，整个系统涉及的功能模块主要有：影视数据爬虫模块、影评分析模块、影视作品分析模块、可视化呈现模块，每个功能模块详细的需求说明如下：

1. 影视数据爬虫模块

如表2-1所示，影视数据爬虫模块的主要功能点和详细介绍如下：

表2-1 影视数据爬虫模块需求分析表

|  |  |  |
| --- | --- | --- |
| 功能模块 | 功能点 | 需求描述 |
| 影视数据爬虫模块 | 影视作品客观信息收集 | 1. 收集影视作品的客观信息，包括电影名称、作品简介、电影类型、主创团队、电影档期、票房数据等 2. 收集影视行业从业人员相关信息，如演员信息、导演信息、编剧信息等 |
| 社交媒体数据收集 | 1. 收集微博、豆瓣等平台上的相关评论信息，针对每部电影收集相应的评论数 |

1. 影视评论分析模块

如表2-2所示，影视评论分析模块的主要功能点和详细介绍如下：

表2-2 影视评论分析模块需求分析表

|  |  |  |
| --- | --- | --- |
| 功能模块 | 功能点 | 需求描述 |
| 影评分析模块 | 影视评论积极、消极情感倾向分析 | 1. 结合表情符号、情感词典和机器学习算法，实现影评两级情感分析 |
| 影视评论积极、中性、消极情感倾向分析 | 1. 结合表情符号、情感词典和机器学习算法，实现影评多级情感分析 |
| 影视评论意见提取 | 1. 实现评论关键词提取，实现影评意见挖掘 |

1. 影视作品分析模块

如表2-3所示，影视作品分析模块的主要功能点和详细介绍如下：

表2-3 影视作品分析模块需求分析表

|  |  |  |
| --- | --- | --- |
| 功能模块 | 功能点 | 需求描述 |
| 影视作品分析模块 | 受众人群分析 | 1. 基于观影人群的性别、收入、兴趣、学历、地域，分析受众人群的分布，定位影片的目标人群 |
| 影人主创指数计算 | 1. 主创人员主要考虑作品的导演及主演，针对主创人员的过往影视作品的票房信息，统计历年作品的总票房 2. 基于历年总票房的增长情况，分析历年作品的权重，基于时间权重与角色权重，计算主创人员票房影响力 |
| 影片主创指数计算 | 1. 基于影人的主创指数，基于角色身份，计算影片的主创指数 |
| 影视票房预测 | 1. 基于影视作品基本信息、影视口碑、主创指数，实现影视票房分档预测 |

1. 影视作品信息可视化呈现

如表2-4所示，可视化呈现模块的主要功能点和详细介绍如下：

表2-4 可视化呈现模块需求分析表

|  |  |  |
| --- | --- | --- |
| 功能模块 | 功能点 | 需求描述 |
| 可视化呈现模块 | 影视信息呈现 | 1. 影视作品及影视演员信息，基于网页的方式，进行可视化呈现，实现简单的电影情报站功能 |
| 影视票房、声量等统计曲线绘制 | 1. 基于影片每日票房数据，绘制影视作品的票房曲线，辅助影视票房的预测分析 2. 收集影视作品的微博声量，绘制相应的声量曲线 |
| 影视评论情感倾向分布绘制 | 1. 基于微博的影视评论，进行情感分析，计算情感分布情况，以及抽取关键词 |
| 影视票房预测曲线绘制 | 1. 基于历史票房数据及电影信息、大众声量，实时预测未来一周的电影票房，并绘制相应的预测曲线，与实际票房曲线开展对比 |

透过以上需求分析，整个系统的用户主要有两种角色：影视作品分析系统和终端用户。关于影视作品分析系统，相应的用例图如图2-1所示，主要涉及影视受众人群分析、影视口碑分析、影片票房预测、影片主创团队影响力分析等用例。

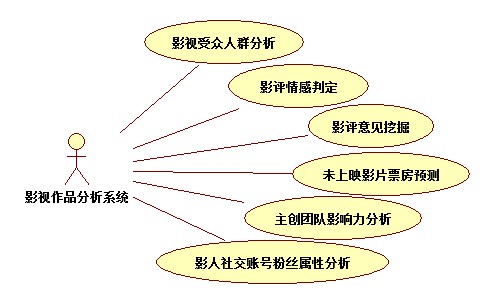


图2-1 影视作品分析系统用例图

基于影视作品分析系统所涉及的用例，以及整个系统最终为终端用户的服务和功能，为终端用户设计的用例如图2-2所示，包括基本的影视信息查看、影人主创指数查询、影片预测票房、受众人群分布、口碑情况等。

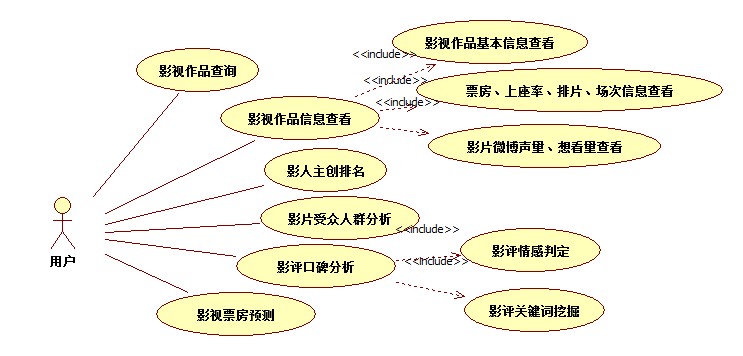


图2-2 终端用户用例图

### 影视票房预测系统非功能需求

本系统的非功能需求主要包括系统性能需求、易用性需求、扩展性需求、正确性需求，健壮性需求等。针对各个非功能需求，具体描述如下：

1. 性能需求

能够及时响应用户的请求，影视作品的相关信息及查询做到快速反馈，影视评论分析及影视价值评估等学习模型在后端开展学习和训练，前端的分析和评估耗时不应超过1s，整个系统操作无卡顿现象。

1. 易用型需求

软件使用者在阅读说明文档后，可以快速上手，学习开销小；系统内的功能操作方法简单，界面简洁，用户能够快速找到自己所需的功能。

1. 扩展性需求

围绕影视数据，可以实现和发散的功能还有很多，因此应保持现有系统接口的简洁和低耦合性，整个系统的接口和功能易于扩展。

1. 正确性需求

本系统要求影视评论的情感极性分析及影视价值评估的准确率维持在一个较高水平。在国内外近期的研究中，针对社交评论的情感分析，特定领域下两级情感判定已达到90%左右准确率的水平，三级情感判定可以达到70%左右水平。因此在项目实现过程中，考虑到分词器及中文情感词典等因素的限制，则要求系统的正逆向两级情感分析准确率维持在80%以上，正逆向中性三级情感分析维持在65%以上的准确率，使整个系统真正具备商业应用价值。

1. 健壮性需求

综合考虑各种情况，保证系统能够正常运行，不会因为突发事件导致数据丢失或者出现差错。对用户的误操作要有一定容忍度，不会因为错误操作而崩溃。同时影视评论分析及影视价值评估具备一定的容错能力，一定程度上过滤噪声语料及错误数据。

## 影视票房预测系统总体设计方案

### 影视票房预测系统架构设计

整个系统的架构设计主要划分为数据获取层、数据存储层、会话层、应用层，具体架构如图2-3所示：

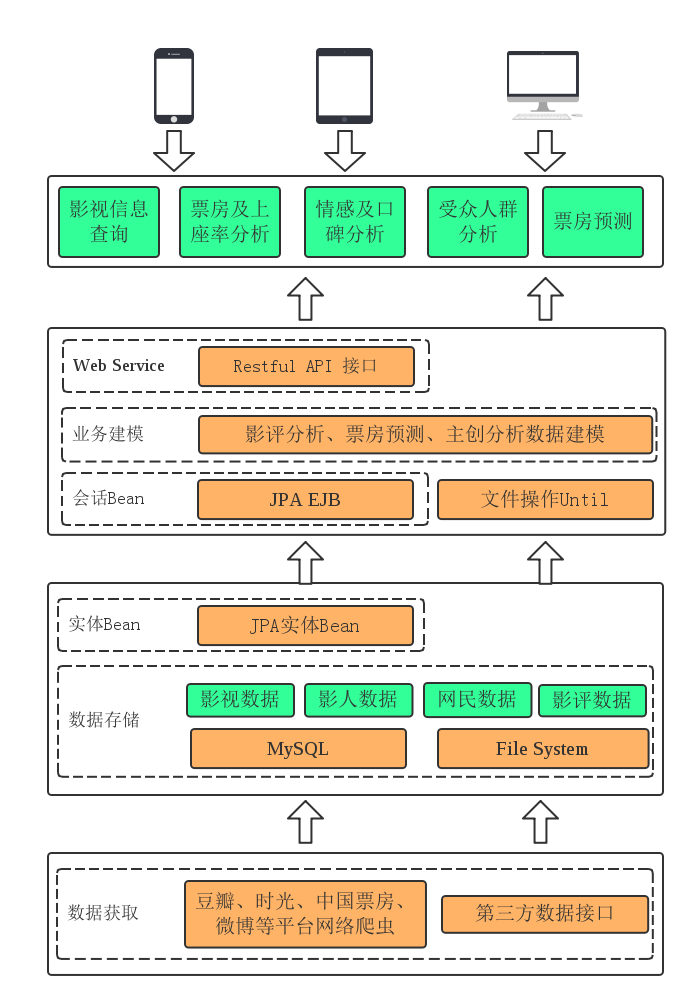


图2-3 系统整体架构

1. 数据获取层

数据获取层主要借助现有的爬虫框架web harvest结合ccrawer框架，同时scrapy作为爬虫辅助，将整个爬虫系统部署到多台服务器上实现，实现高并发的数据爬取和收集。

第三方接口，则主要应用微博提供的数据接口，由于爬虫的噪声较大，所以针对实时产生大量数据的微博内容，则采用购买数据接口的方式，收集相关数据。

1. 数据存储层

数据存储则主要基于关系型数据库MySQL，同时辅助文件系统存储项目所需的模型文件，以及分词器词典。与此同时，基于数据库的表单设计，反向 生成实体Bean，借助JPA的方式操作数据库，简化了SQL语句的书写及持久层的维护。

1. 会话层

会话层主要负责系统核心业务逻辑的书写，一方面基于会话Bean，封装底层数据库的操作。另一方面，借助IKAnalyzer、weka等开源工具，实现情感分析模型以及票房预测模型的集成。接下来借助servlet以及jersey框架，封装成restful API，以便应用层页面的调用。

1. 应用层

应用层主要基于JS异步调用的方式，借助restful接口，实现业务功能的调用。同时，借助bootstrap框架，实现界面的响应式布局，使其在不同设备上都能有较好的适配。

### 影视票房预测系统功能结构设计

影视票房预测系统模块划分及功能结构，如图2-4所示，涉及影视作品数据爬取、影视评论分析、影视作品分析、影视作品信息呈现四个模块，每个模块的简要说明如下：

1. 影视数据爬取模块

主要针对国内猫眼、微票、时光网、中国票房、新浪微博等权威网站，进行数据爬取，以及对数据的整合和去噪处理，进行数据存储，支撑后续的信息呈现和分析任务。

1. 影视评论分析模块

主要从微薄影评出发，基于文本挖掘的方式，分析评论的情感倾向，以及提取评论中的关键词，分析大众意见。

1. 影视作品分析

主要从受众人群的分布、主创团队的影响力、影视票房的预测三个角度展开，实现影视作品的自动化分析，从中挖掘有价值的知识及信息。

1. 影视作品信息呈现

做到影视领域的搜索引擎，具备完整而详细的影视信息，以及时刻掌握最新的票房、排片、声量、评论等动态，完整的呈现一部影视作品的各个方面的信息。



图2-4 系统功能结构图

## 本章小结

本章对项目中将开发的系统进行了需求分析，确认了系统的功能需求和非功能需求，并基于需求分析的结果进行了系统架构设计、模块划分、功能结构设计。为接下来的系统详细设计以及实现奠定了基础。

# 影视票房预测系统详细设计

本章在上一章节需求分析与总体设计的基础上，对系统进行了详细设计。主要设计了系统所用到的数据结构、核心类的详细类图以及系统主要算 法。

## 影视票房预测系统存储设计

整个系统在具体数据结构设计中，主要有两大数据存储模块：社交平台信息存储和影视作品信息存储。社交平台信息存储则包括来自不同社交平台的评论信息及账号信息，从而支撑影评情感分析及影片受众人群分析。影视作品信息，则主要包括影视作品的基本信息、影人信息、预告片信息、票房信息等。

### 社交影评及用户信息存储设计

如图3-1所示，社交影评及用户信息主要涉及微博、豆瓣等影评信息和账号信息，涉及的表单有vperson, netizen, netizen\_status, netizen\_property, emotion，表与表之间主要为一对多及一对一的关联关系。

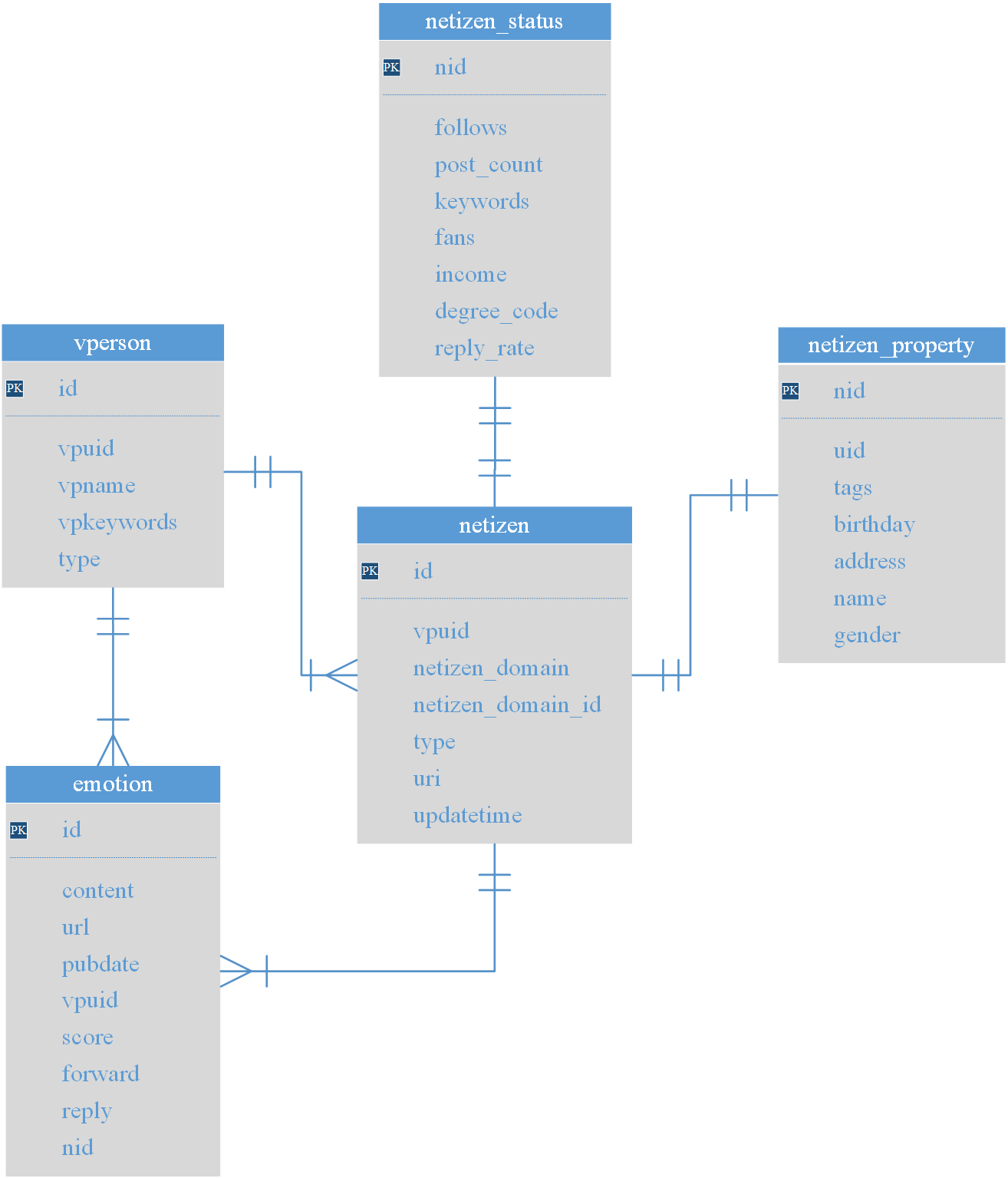


图3-1 社交影评及用户信息存储设计

其中，vperson表记录爬虫任务的相关信息，每一次关于影片信息爬虫任务的发起，都会在vperson表中插入数据，具体字段说明如表3-1所示：

表3-1 vperson表字段说明

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段类型 | 字段含义 |
| vpuid | bigint (20) | 任务id |
| vpname | varchar(255) | 任务信息描述 |
| vpkeywords | varchar(255) | 爬虫任务关键词 |
| type | smallint(6) | 任务类型 |

关于netizen表，则主要是基于爬取到的评论，同时收集社交账号信息，详细字段说明如表3-2所示：

表3-2 netizen表字段说明

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段类型 | 字段含义 |
| id | bigint (20) | 爬取到的网民自增id |
| vpuid | bigint (20) | 对于爬虫的任务id |
| netizen\_domain | varchar(100) | 社交账号平台 |
| netizen\_domain\_id | varchar(255) | 社交账号id |
| type | smallint(6) | 类型 |
| uri | varchar(255) | 社交账号主页网址 |

关于netizen\_property表，则记录社交账号上用户的个人信息，辅助受众人群分析，详细字段说明如表3-3所示：

表3-3 netizen\_property表字段说明

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段类型 | 字段含义 |
| nid | bigint (20) | 对应netizen表id |
| uid | varchar(255) | 社交账号id |
| tags | varchar(255) | 用户标签信息 |
| birthday | date | 生日 |
| name | varchar(50) | 真实姓名 |
| gender | varchar(30) | 性别 |

关于netizen\_status，如表3-4所示，主要记录用户的活动情况及动态信息。

表3-4 netizen\_status表字段说明

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段类型 | 字段含义 |
| nid | bigint (20) | 对应netizen表id |
| follows | int(11) | 关注量 |
| post\_count | int(11) | 发帖量 |
| fans | int(11) | 粉丝量 |
| income | int(11) | 收入水平 |
| degree\_code | int(11) | 学历情况 |

关于emotion表，则主要记录从社交平台爬取下来的评论，并存储情感判定结果，字段详细说明如表3-5所示：

表3-5 emotion表字段说明

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段类型 | 字段含义 |
| id | int(11) | 自增id |
| content | text | 评论内容 |
| url | varchar(256) | 评论URL |
| pubdate | datetime | 发布时间 |
| forward | int(11) | 转发量 |
| reply | int(11) | 回复量 |
| nid | bigint(20) | 发表评论用户id，对应netizen表的id |
| score | int(11) | 情感倾向 |

### 影视作品信息存储设计

影视作品信息存储主要涉及影片、影人及票房信息，相应的信息存储设计如图3-2所示：

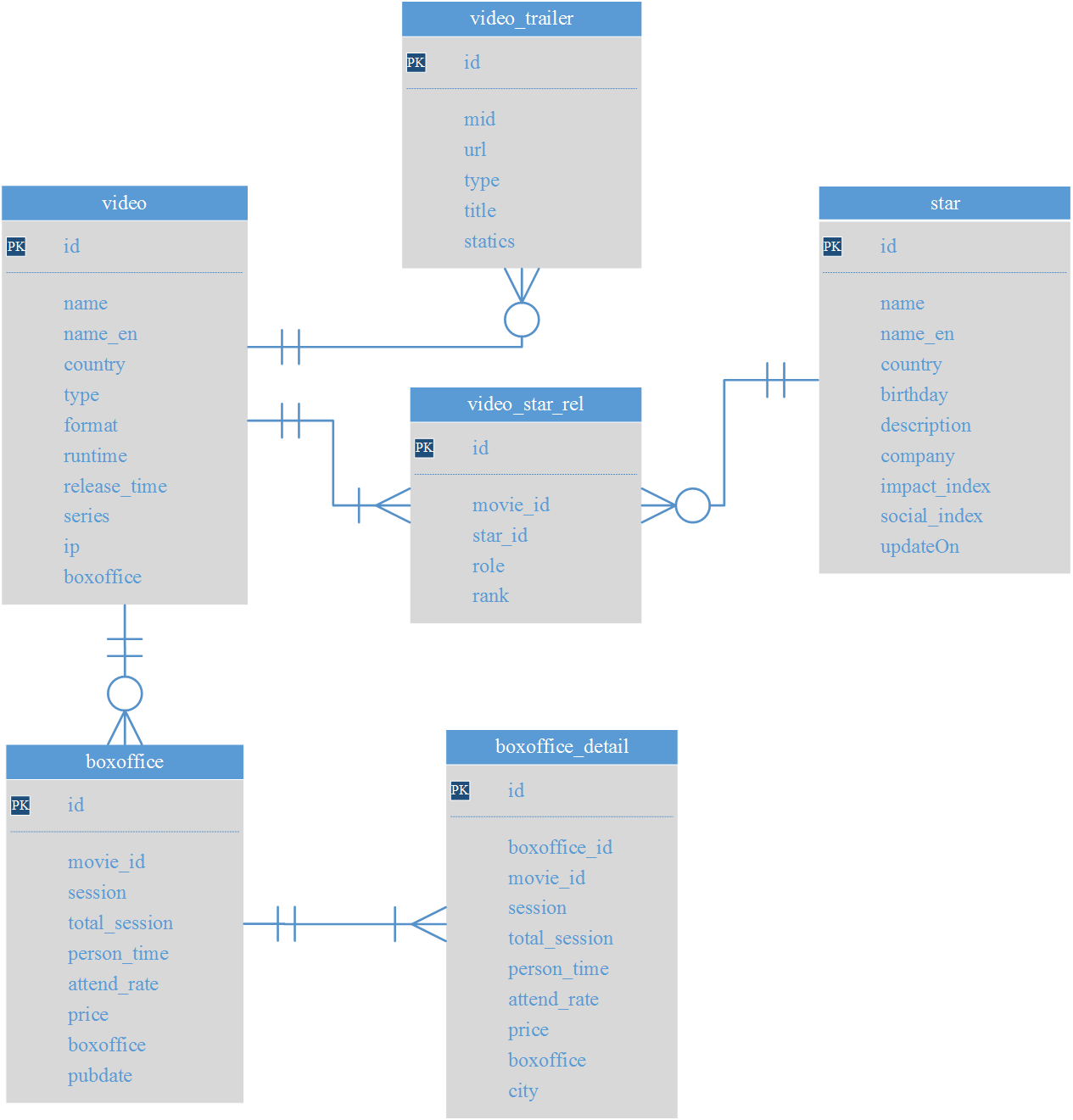


图3-2 影视作品信息存储设计

影视信息核心数据表video，记录影片的相关信息以及票房信息，相关字段说明如表3-6所示：

表3-6 video表字段说明

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段类型 | 字段含义 |
| id | int(11) | 影视id，自增长 |
| vpuid | bigint(20) | 影视信息对应的爬虫任务id |
| name | varchar(255) | 影视作品名 |
| name\_en | varchar(255) | 影视作品英文名 |
| type | varchar(255) | 影视类型 |
| country | varchar(255) | 发行地区 |
| runtime | varchar(255) | 时长 |
| format | varchar(255) | 影片制式，2D/3D/IMAX |
| releasetime | datetime | 上映时间 |
| byname | varchar(1024) | 影片别名 |
| introduction | text | 影视介绍 |
| series | int(11) | 是否系列电影 |
| ip | int(11) | 是否ip电影 |
| boxoffice | int(11) | 电影票房 |
| predict\_boxoffice | int(11) | 预测票房 |

影人信息核心数据表star，记录影人的基本信息及社交账号相关字段说明如表3-7所示：

表3-7 star表字段说明

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段类型 | 字段含义 |
| id | int(11) | 影人id，自增长 |
| vpuid | bigint(20) | 明星对应的vpuid，关联声量使用 |
| name | varchar(255) | 明星姓名 |
| name\_en | varchar(255) | 明星英文名 |
| birthday | datetime | 生日 |
| gender | varchar(1024) | 性别 |
| homeplace | varchar(1024) | 出生地 |
| socialmedia | varchar(1024) | 微博url |
| description | varchar(1024) | 简介 |
| impact\_index | double | 演员票房影响力(运算生成) |
| social\_index | double | 演员社交影响力(运算生成) |
| updateOn | datetime | 更新时间 |

影视信息与影人的关联表video\_star\_rel，每一条影片对应多条记录，影人所承担的角色也记录在该表中，字段说明如表3-8所示：

表3-8 video\_star\_rel表字段说明

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段类型 | 字段含义 |
| id | int(11) | 主键，自增长 |
| movie\_id | int(11) | video表对应id，关联影片 |
| star\_id | int(11) | star表对应id，关联影人 |
| role | tinyint(3) | 影人在该部影片中所扮演的身份 |
| rank | int(11) | 影人角色排名 |

影视票房信息数据表boxoffice，记录排片、上座率等相关信息，字段说明如表3-9所示：

表3-9 boxoffice表字段说明

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段类型 | 字段含义 |
| id | int(11) | 主键，自增 |
| movie\_id | int(11) | 关联的影片id |
| session | int(11) | 统计场次 |
| total\_session | int(11) | 总场次 |
| person\_time | int(11) | 人次 |
| attendance\_rate | varchar(255) | 上座率 |
| price | smallint(6) | 票价 |
| boxoffice | int(11) | 日票房 |
| pubdate | datetime | 日期 |

影视票房地域详细信息数据表boxoffice\_detail，记录不同城市排片、上座率等相关信息，字段说明如表3-10所示：

表3-10 boxoffice\_detail表字段说明

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段类型 | 字段含义 |
| id | int(11) | 主键 自增 |
| boxoffice\_id | int(11) | 对应boxoffice表id |
| session | int(11) | 统计场次 |
| total\_session | int(11) | 总场次 |
| person\_time | int(11) | 人次 |
| attendance\_rate | varchar(255) | 上座率 |
| price | smallint(6) | 票价 |
| boxoffice | int(11) | 日票房 |
| city | varchar(255) | 城市 |

影视作品预告片信息数据表video\_trailer，记录影片在不同视频平台下的预告片播放情况，相关字段说明如表3-11所示：

表3-11 video表字段说明

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段类型 | 字段含义 |
| id | bigint(20) | 自增id |
| mid | bigint(20) | 关联影片id |
| url | varchar(1024) | 预告片连接地址 |
| type | varchar(255) | 类型 |
| title | varchar(1024) | 预告片全名 |
| statics | mediumtext | Json格式播放量统计信息 |

## 影视票房预测系统核心算法设计

### 影评情感分析算法设计

如图3-3所示，影视评论情感分析算法实现整体流程设计如下，涉及数据收集、中文分词、特征选取、机器学习等相关工作。

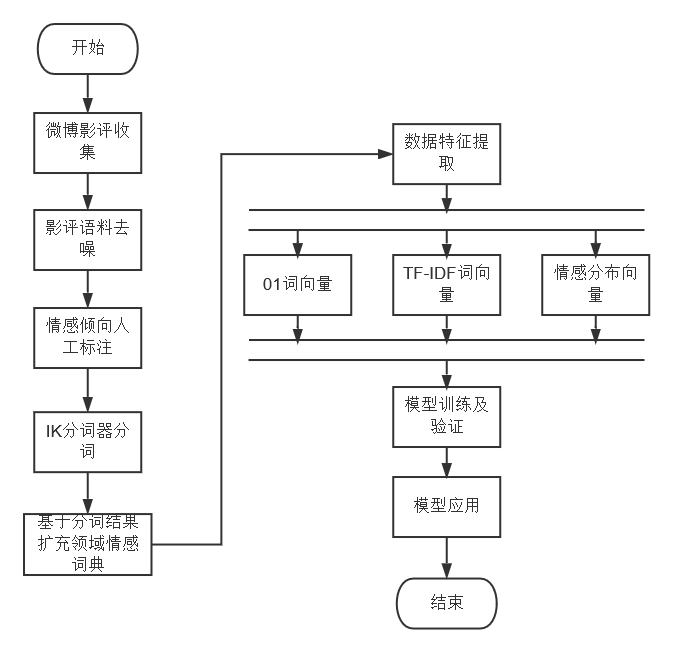


图3-3 影视评论情感分析流程图

1. 数据收集

影评情感倾向分析的训练语料主要来自新浪微博上的影评信息，具体开展过程中，收集数据库中2013~2015年的电影评论，累计记录40w+，为了防止大量评论来自同一部电影，从中随机抽取1w条语料作为初始训练数据。

1. 数据去噪

由于微博属于大众媒体平台，用户群体中也会存在部分非真实用户，常常发布宣传与广告性的相关微博，同时由于微博的社交因素，评论语料中也会存在许多@交互的内容。因此，为了提升影视评论情感分析的准确性，对于初始语料，需要开展以下工作：

* 过滤广告宣传性微博
* 删除微博评论中的URL
* 删除用户@其他用户的文本内容
* 提取微博中的表情信息(中括号中文本)，以便后续分析使用

1. 数据标注

数据标注过程中，需要将影视评论最终划分为中性、积极、消极三种情感倾向。其中，正向/积极评论，标注为1；逆向/消极评论，标注为-1；中性评论(感情色彩并不强烈)，标注为0。同时在数据标注过程中，同时开展数据过滤的工作，对于相关性较差的评论进行人工过滤。过滤规则如下：

* 语料中存在大量的符号或者乱码
* 评论内容为电影的官方宣传或公告信息
* 评论内容与影视相关度较低
* 微博评论中相互@转发回复的内容

1. 中文分词

采用IKAnalyzer及Hanlp分词器开展中文分词，具体在分词过程中，引入命名实体识别，针对文本出现的人名、地名、组织名进行识别和处理，提高分词的准确度。分词后的语料样例如图3-4所示：

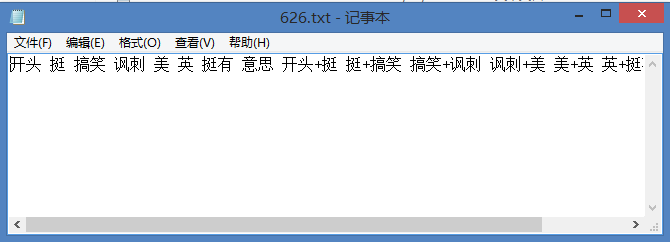


图3-4 分词后语料样例

1. 核心情感词典扩充

在许多文本情感分析的文章中，都有提及情感词典。结合情感词典特征，一方面可以从语义角度分析句子的情感倾向，另一方面也为机器学习算法提供有力的数据特征。在个人实施过程中，借助了中文情感词典有大连理工的DUTSD(2008)，台湾大学的NTUSD(2006)，但最终发现，结合DUTSD以及基于分词语料整理的领域情感词效果最好，借助情感词典，在后续的分类过程中，辅助情感分类的初步划分，具体构造的情感词典如表3-12所示，标注情感倾向以及倾向程度。

表3-12 情感词典样例表

|  |  |  |
| --- | --- | --- |
| 情感倾向 | 倾向度 | 样例 |
| 积极 | 4 | [震撼,4.0] [非常好,4.0]  [美好,4.0] [唯美,4.0] |
| 3 | [有趣,3.0] [感动,3.0]  [很美,3.0] [可爱,3.0] |
| 2 | [福利,2.0] [挺好,2.0]  [佩服,2.0] [强大,2.0] |
| 1 | [贡献,1.0] [珍惜,1.0]  [坚守,1.0] [风采,1.0] |
| 消极 | -4 | [糟糕,-4.0] [无耻,-4.0]  [垃圾,-4.0] [恶心,-4.0] |
| -3 | [浪费,-3.0] [幼稚,-3.0]  [不好,-3.0] [放屁,-3.0] |
| -2 | [曝光,-2.0] [自私,-2.0]  [不懂,-2.0] [鄙视,-2.0] |
| -1 | [可怕,-1.0] [勉强,-1.0]  [害怕,-1.0] [娱乐,-1.0] |

1. 数据特征选择

由于整个文本情感分类模型，主要基于词袋模型，则在具体词向量特征的选择上，主要考虑单词元以及双词元，最终实验过程中，发现单词元结合双词元的效果最好。

在词向量的特征转换上，也主要尝试了以下三个方面：

a) 01词向量，将每条文本映射成分词的01向量，1表示文本中存在该分词；

b) TF-IDF向量，统计训练语料中每个词的TF-IDF值，将文本转换为对应分词的TF-IDF向量；

c) 情感分布向量，针对每个词，根据公式(3-1)，统计在训练语料中不同情感分类中的概率分布，其中第个词在第个分类的概率分布为：

(3-1)

其中：

* + 为第*i*个词在第*j*类中的词频计数；
  + 为第*i*个词的总词频统计；
  + 为第*i*个词第*j*类的概率值；

整篇文本的特征向量，如公式(3-2)，为每个词的情感分布向量累和获得，记为：

(3-2)

其中：

* + 为第*d*篇文档第*j*类的特征值；
  + *n*为该篇文档的分词个数；

1. 模型训练和验证

模型选择上，主要采纳SVM、Naïve Bayes、Max Ent三类方法，同时针对SVM，尝试不同kernel的变换。

在算法模型的验证上，除了交叉验证，还主要考虑采用划分训练集和测试集的方式，采用7:3的比例，多次随机选择训练集和测试集，统计模型的准确率，并且结合precision, recall, F1 score，最终分析算法的性能和好坏。

1. 模型实现及应用

影评情感分析的核心模型具体实现则主要基于java代码，核心机器学习模型，则采用源码的方式调用机器学习工具包weka。

### 影评关键词提取算法设计

谈起关键词提取算法，常见的并且最易实现的当属TF-IDF，但是对于微博影评这样的短文本，TF-IDF效果一般，因此采纳TextRank算法。

TextRank是在Google的PageRank算法启发下，针对文本里的句子设计的权重算法，目标是自动摘要和关键词提取。它利用投票的原理，让每一个单词给它的邻居（术语称窗口）投赞成票，票的权重取决于自己的票数。

首先，PageRank的算法如公式(3-3)所示：

(3-3)

而TextRank算法则是在PageRank的公式的基础上，引入了边的权值的概念，代表两个分词或者分句的相似度。具体算法如公式(3-4)所示：

(3-4)

其中：

a) 是第*i*个单词的权值；

b) d为阻尼系数，常数；

c) 、表示指向第i个单词的集合和第i个单词指向的集合，在TextRank算法中，两个集合可以相同，在特定窗口距离之间的词都在集合之内；

d) 初始时，可以设置每个词的权重为1，经过多伦迭代后，词权重稳定后，按照词权重的大小排序，即可提取关键词；

具体在系统应用时，基于TextRank的影评关键词算法流程则如图3-5所示：

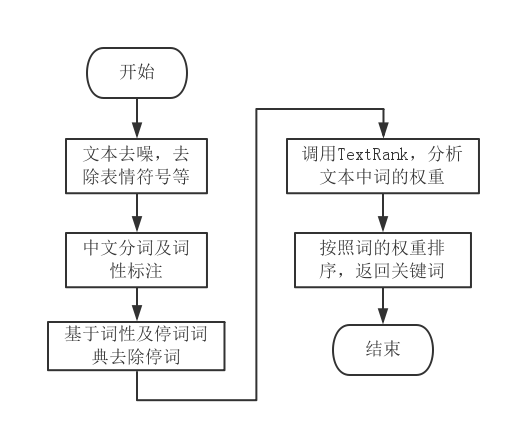


图3-5 影评关键词提取算法流程设计

### 主创指数算法设计

1. 主创票房矩阵

每一位主创人员基于近五年以不同身份参与电影的作品票房，都可以得到一个的矩阵。第位主创的矩阵为，其中表示第位主创人员在阶段以第职业的身份参与电影的总票房，表示职业为导演，表示职业为主演，表示职业为编剧，表示职业为制片，那么有：

 (3-5)

1. 阶段权重

主创人员的票房影响力主要考虑导演及主演两大主创人群。通常一位主创在各个阶段参与的电影，对其将要上映的电影的影响是不同的，一般是时间越靠近将要上映的电影，其影响力越强。所以，针对每一位主创人员，基于近五年作品的票房情况，划分成，第一个阶段(最近一年2015年票房)时间内作品的总票房最具有代表性的，向前推，从2015年到2011年其代表性依次降低。同时，最近十几年，中国大陆电影市场每年的总票房在呈指数趋势增长。当下大陆市场容量在急速扩充，后期上映的电影，其能够吸纳的票房肯定高于前期上映的电影。所以基于代表性和票房发展趋势这两点因素的考虑，我们采用如下方法对各个阶段赋权。

以当前时间点2016年，近五年的年度票房及作品数如表3-13所示：

表3-13 年度票房历史数据

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 年份 | 2015 | 2014 | 2013 | 2012 | 2011 |
| 阶段 |  |  |  |  |  |
| 票房 | 441 | 296 | 218 | 171 | 131 |
| 上映影片数 | 358 | 323 | 305 | 289 | 228 |

如公式(3-6)，各个阶段的平均票房为：

(3-6)

如公式(3-7)，五年内累计平均票房为：

(3-7)

如公式(3-8)，基于平均票房计算各个阶段的权重为：

(3-8)

1. 角色权重

在数据中，我们将电影主创人员划分成了以下四种职业，分别是导演()，主演()，编剧()，制片()。因为每类职业对电影最终票房的贡献不一样，所以对每类主创人员的赋权也不一样。

从观众和投资方的角度，以及当下电影市场的销售状况，凭借主观经验可以判断，对电影票房影响最大的首先是电影的主演整容，一部电影的主演整容越强大，社会曝光率越高，吸引观众的能力就越大。然后是导演，导演的水平直接关系着一部电影的质量。之后是编剧，最后是制片。制片是幕后创作者，很少有人会因为一部电影的制片去观看一部电影的，但是制片在电影生产环节中的好坏，会直接影响这电影品质和营销的质量，从而间接影响票房。所以如公式(3-9)，这里我们对导演()，主演 ()，编剧()，制片()依次赋权为：

(3-9)

虽然这种方法是通过主观的方法获得的，但是就目前计算的结果来看效果是最好的。通过相关性分析，这种方法最终获得主创人员得分与被预测的电影票房数据相关性较高。

1. 主创指数

根据上面得到的票房矩阵和权重向量，基于公式(3-10)，我们可以计算出第 位主创的打分，并以此衡量主创人员的票房影响力，辅助影视票房的预测：

(3-10)

### 受众人群分析算法设计

影片的受众人群信息，主要来自于社交平台上的网民信息，基于网民发表的影片评论，收集网民信息，基于网民的基本信息和标签分析，从年龄、性别、收入、学历、兴趣等角度，进行影片受众群体分析。在受众人群划分中，比较难实现的收入群体分析。因此，在这里提出了网民年收入计算模型，具体算法设计如图3-6所示：

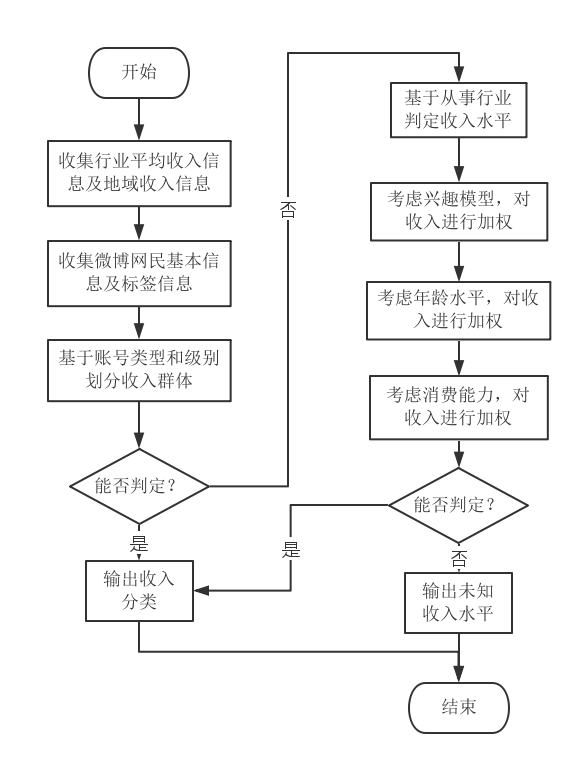


图3-6 网民年收入判定模型

而关于受众人群的学历以及兴趣信息则基于用户标签和基本信息进行获取，算法流程如图3-7所示：

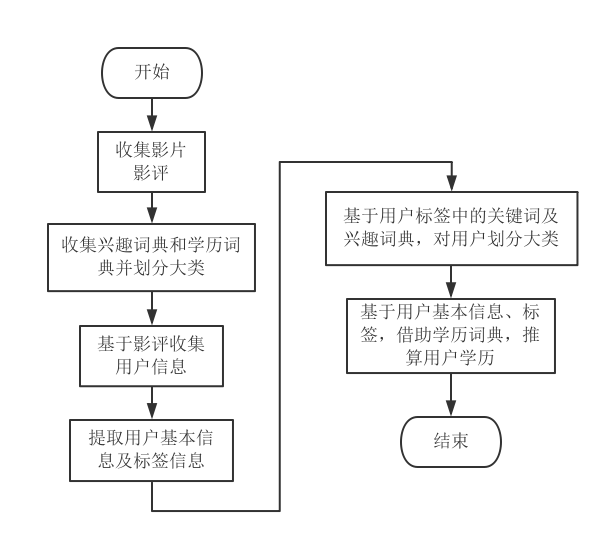


图3-7 网民学历及兴趣爱好判定模型

### 影视票房预测算法设计

影视票房预测的整体框架如图3-8所示：整个流程中，涉及到的内容有影视作品基本信息、社交评论的数量特征、情感特征、主创指数、预测模型等。

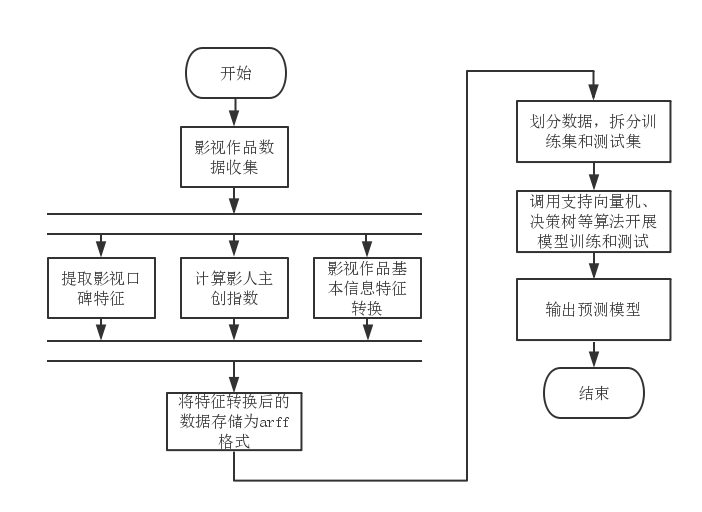


图3-8 影视票房预测整体流程图

影视作品口碑特征提取如图3-9所示，一方面基于影评数量，提取口碑数量特征；另一方面，基于影评情感分析，分析影片的口碑倾向，将不同情感倾向的倾向度作为数据维度，融入到影片票房预测中。

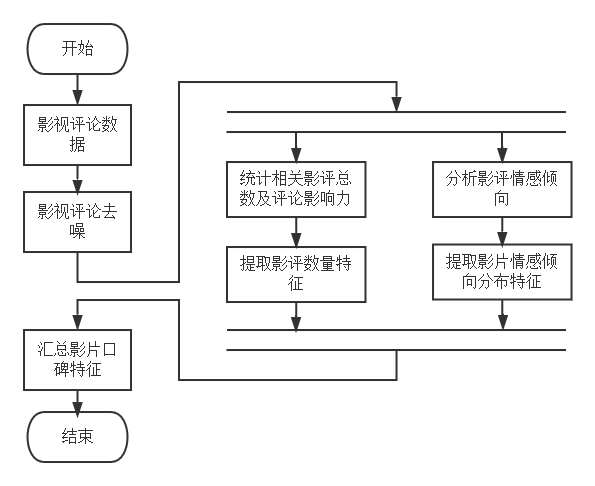


图3-9 影片口碑特征提取流程图

整个票房预测算法，主要基于机器学习的方式，提取影片不同的数据维度，进行模型训练，最终得出票房预测模型，其中，考虑的数据维度有：

* 导演、演员票房影响力
* 导演、演员社交影响力，如粉丝量
* 影视类别(0：爱情 1：动作 2：喜剧 3：剧情 4：科幻 5：魔幻 6：动画 7：惊悚 8：战争 9：纪实 10：歌舞)
* 档期(依据上映时间，推算档期，暑期档，春节档等)
* 影片发行地区(港台、大陆、美国、其他海外)
* 影片制式，是否支持3D，支持IMAX
* 预告片播放量
* 系列电影
* IP电影
* 影评情感倾向及数量特征

影视票房在预测过程中，将其转化为分类问题，将电影票房划分为百万以下、百万、千万、亿、十亿五类作品，基于以上数据维度，开展模型的训练和验证，在具体训练过程中，采用决策树、支持向量机、随机森林、朴素贝叶斯等算法进行尝试和调优，最终挑选出效果最好的三类算法模型，并进行投票得出最终票房预测结果。其中，训练数据采用2011~2015年的影视作品信息，以2016年上映的影片进行模型验证和测试，统计模型的准确率。

## 影视票房预测系统功能实现详细设计

### 影评情感分析实现详细设计

基于影评情感分析算法流程的设计，关于影评情感分析的核心数据结构设计如图3-10所示：

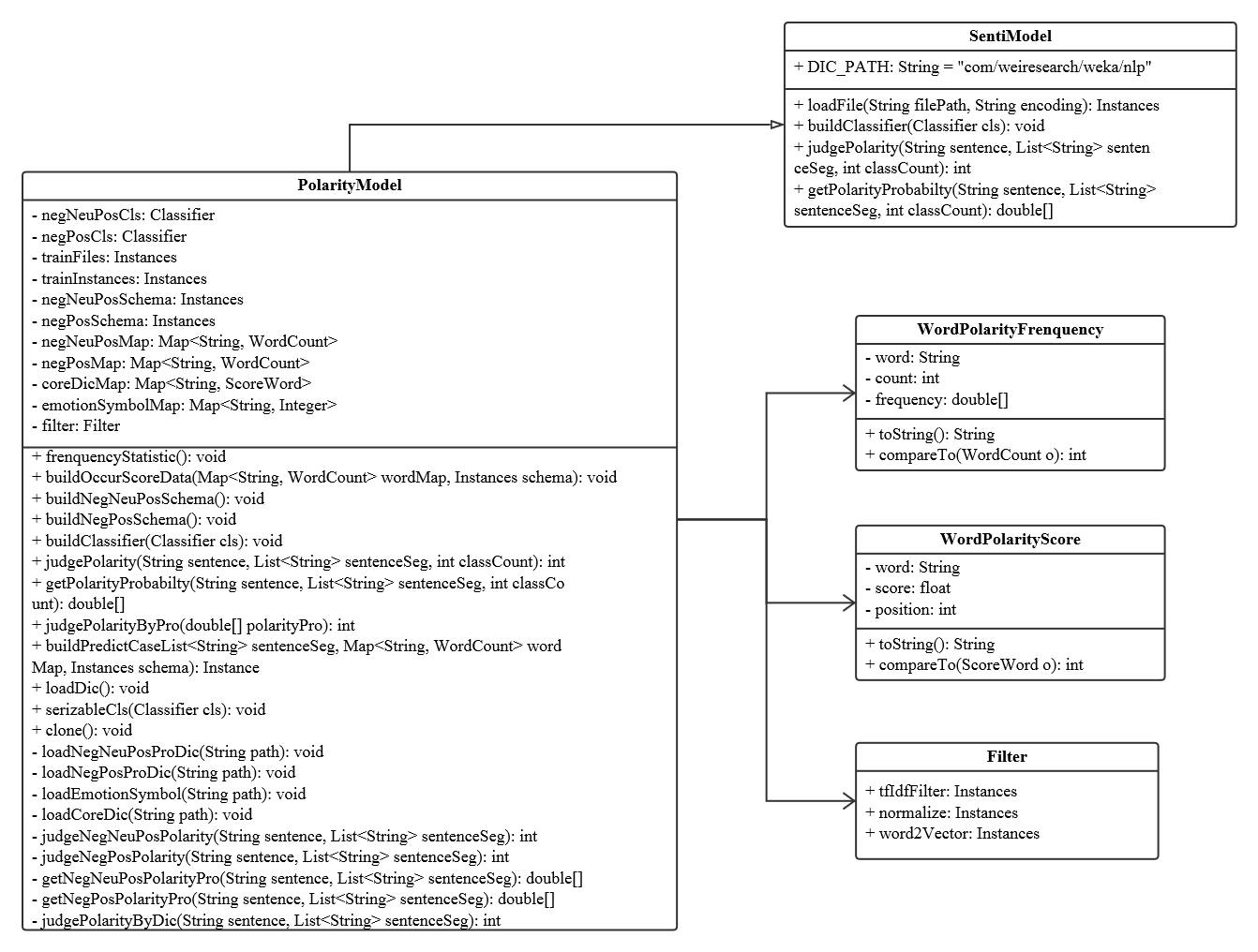


图3-10 影评情感分析类图设计

说明如下：

a) PolarityModel：封装了情感词典的加载、模型的训练、模型的序列化，以及预测接口的暴露；

b) SentiModel：抽象类，封装基本的训练及分类接口，以便不同模型的实现；

c) WordPolarityFrenquency：统计训练语料集中每个词的情感分布向量；

d) WordPolarityScore：记录每个词的情感倾向及倾向度；

1. Filter：实现tf-idf向量、01向量、规范化等数据变换filter；

在核心算法模型之上，由于整个情感分析依赖于分词器，并且最终会封装到web service中，整体架构的数据结构设计如图3-11所示：

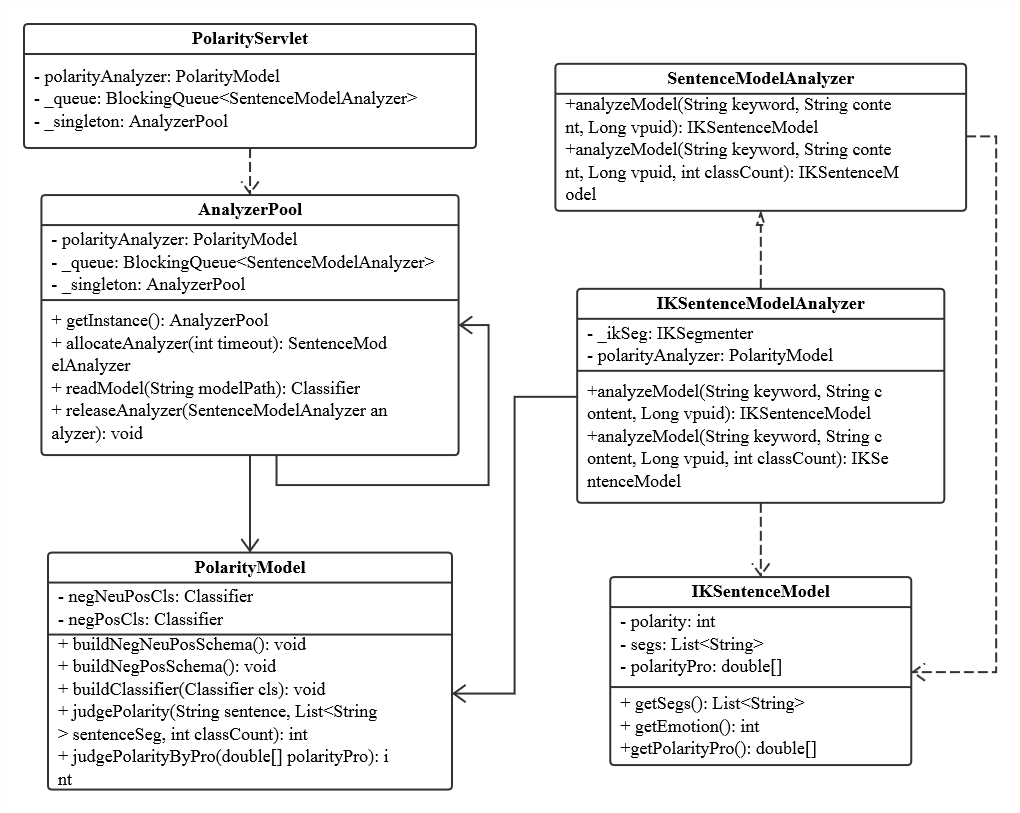


图3-11 影评情感分析整体类图设计

说明如下：

1. PolarityServlet：影评情感分析servlet接口，对外暴露成web service；

b) AnalyzerPool：影评分析线程池类，单例模式设计，每一次分析都从线程池中调度一个线程进行分析；

c) PolarityModel：封装了情感词典的加载、模型的训练、模型的序列化，以及预测接口的暴露；

d) SentenceModelAnalyzer：影评分析器接口，可以基于不同的分词器进行泛化；

e) IKSentenceModelAnalyzer：基于IKAnalyzer分词器的影评分析器实现，进行分词、词性标定、情感标定等工作；

f) IKSentenceModel：影评分析结果实体类，存储分词结果及情感极性判定结果；

在实际调用过程中，整个影评情感分析接口调用的时序图如图3-12所示，Web Server容器启动时，初始化分析器线程池，具体调用时，servlet最先接收分析请求，调度影评分析线程池AnalyzerPool，分配分析器analyzer，进行影评分词和情感分析。

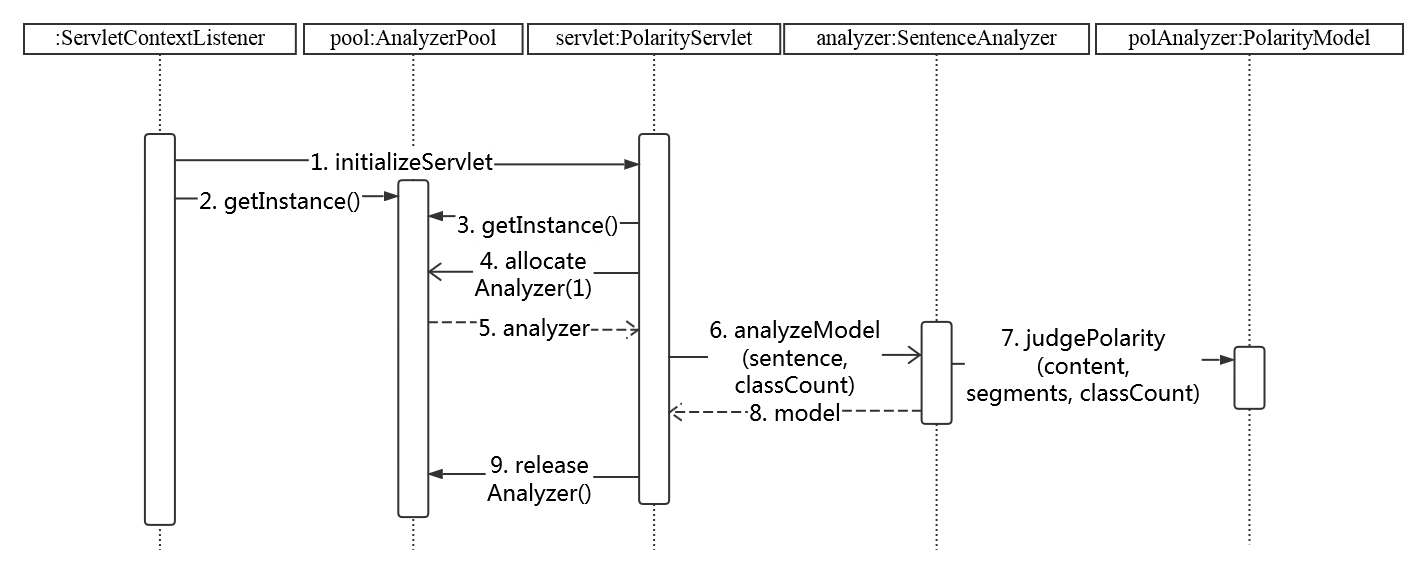


图3-12 影评情感分析时序图设计

### 影人主创指数实现详细设计

基于影人主创指数的计算方法，影人主创指数核心类图设计如图3-13所示：

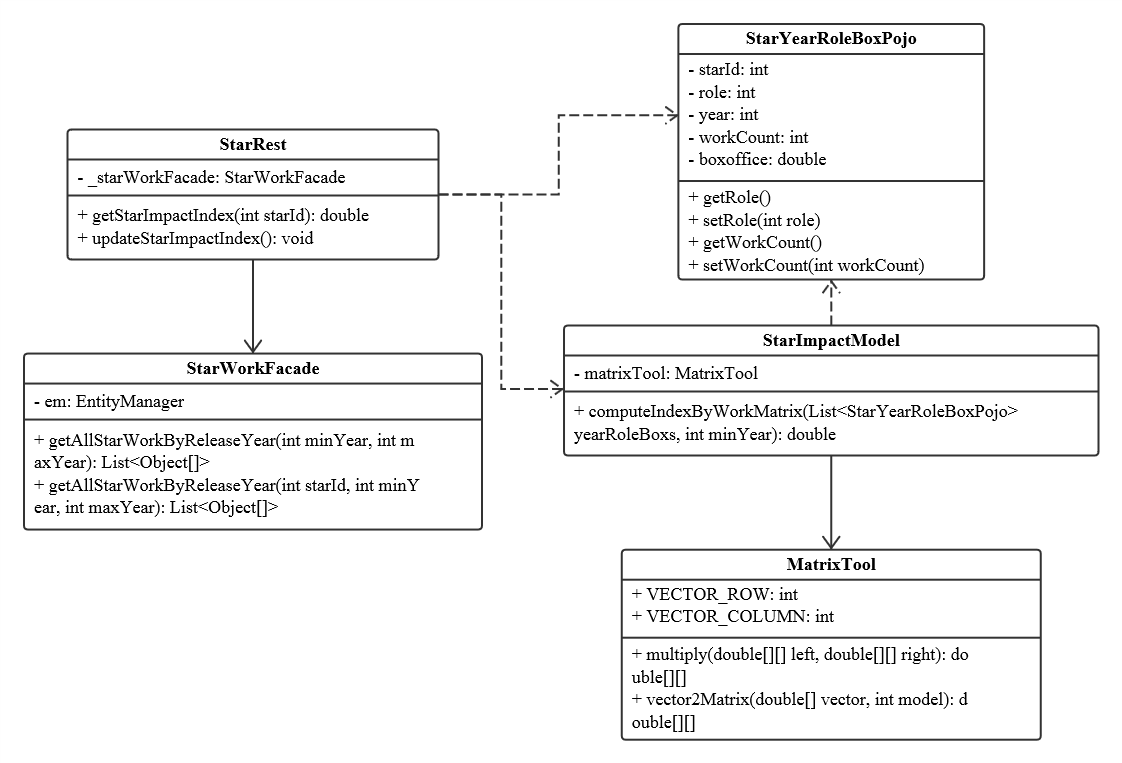


图3-13 影人主创指数类图设计

说明如下：

1. StarRest：影人信息Rest接口，提供影人信息及主创指数查询及更新；
2. StarWorkFacade：影人作品信息控制类，负责影人作品信息增删改查 ；
3. StarYearRoleBox：影人不同年份、角色参与影片获取票房POJO类；
4. StarImpactModel：影人主创指数计算核心实现；
5. MatrixTool：矩阵运算工具类，辅助主创指数计算；

在查询具体主创指数时，时序逻辑如图3-14所示：

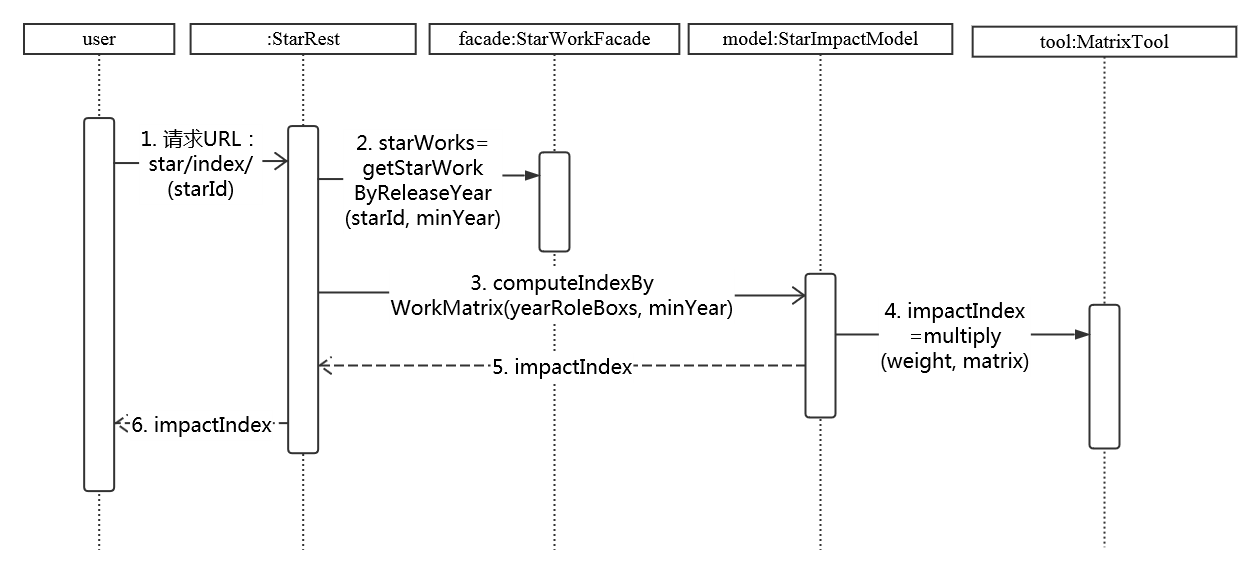


图3-14 影人主创指数查询时序图设计

说明如下：

a) StarRest向用户提供接口，用户请求查询特定演员的主创指数；

b) 调用StarWorkFacade，查询演员作品信息；

c) 基于演员作品信息，统计演员历年以不同角色参与电影获得的票房，代入StarImpactModel进行计算；

d) 基于作品矩阵及权重信息，最终得出主创指数，返回；

### 影视票房预测实现详细设计

影视票房预测依托于影视作品基本信息、影人主创指数、预告片信息、影视口碑信息等，详细类图设计如图3-15所示，说明如下：

a) StarImpactModel：主要用于分析主创人员的影响力指数，提供多种指数计算方法，并最终选择最优的方式，集成到系统中；

b) MovieModel：主要用于数值化影视信息，以及匹配影人信息，生成训练case；

c) BoxPredictModel：影视票房预测核心实现，封装不同的训练方法，及数据特征变换方式；

d) Star：影人信息类

e) StarSocial：影人社交账号信息类，主要记录新浪微博社交账号的信息，以及社交影响力指数；

f) StarImpact：记录主创人员票房影响力和过往作品信息；

g) Moive：记录影视作品的相关信息；

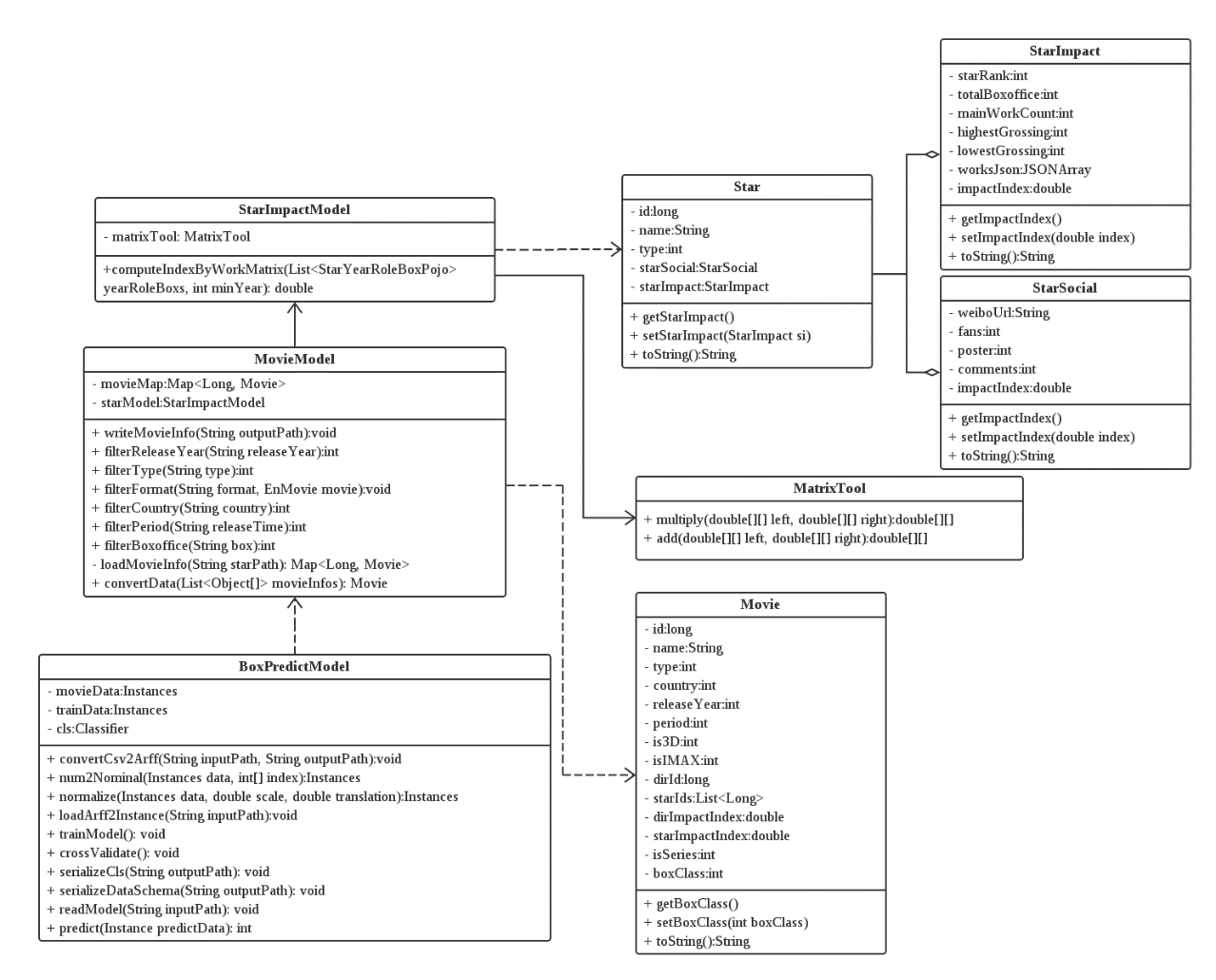


图3-15 影视票房预测类图设计

影视预测接口，训练完成后，还需要进一步集成到web service中，票房预测功能集成详细类图设计如图3-16所示，说明如下：

a) VideoRest：封装与影视信息相关的web service接口；

b) VideoFacade：影视信息数据库Controller，负责影视表单的CRUD操作；

c) SingletonPredictModel：集成多个影视票房预测模型，单例模式实现，具体预测时，多个模型进行投票选出最终结果；

d) BoxPredictModel：影视票房预测核心实现，封装不同的训练方法，及数据特征变换方式；

e) MovieModel：主要用于数值化影视信息，以及匹配影人信息，生成训练case；

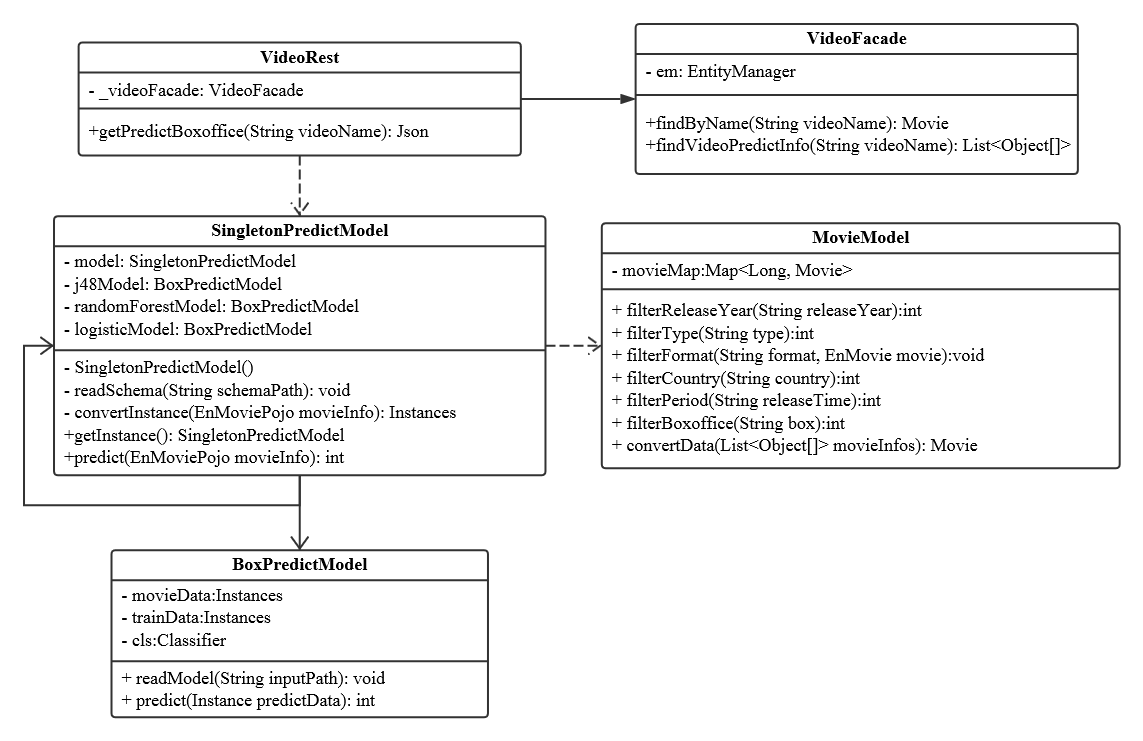


图3-16 影视票房预测集成类图设计

影视票房预测接口基于servlet暴露成web service，用户调用的时序图如图3-17所示，调用流程如下：

1. VideoRest向用户提供接口，用户请求查询特定影片的预测票房；
2. 调用VideoFacade，查询影视作品信息；
3. 基于影视作品信息，查询相关联主创团队的主创指数及影片口碑信息；
4. 综合影视作品各个方面的信息，并进行数据特征转换；
5. 将转换后的数据传给预测接口，进行票房预测并返回；

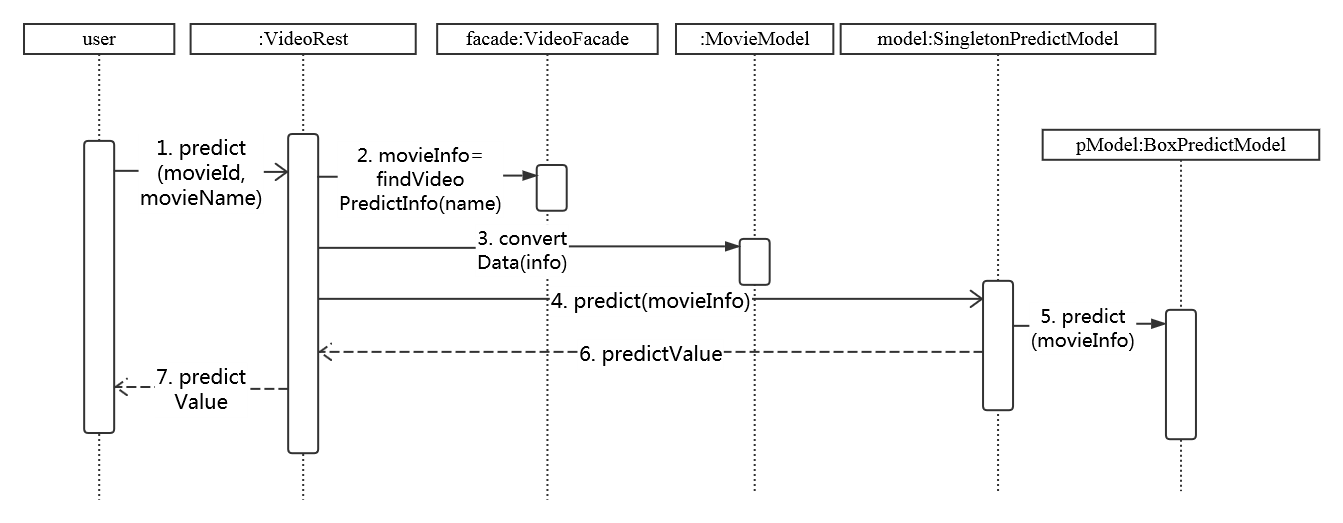


图3-17 影视票房预测接口调用时序图

## 本章小结

本章在总体设计的基础上，对系统进行了详细设计，具体内容包括系统的核心算法流程设计、类的详细设计以及功能模块调用过程中的时序逻辑。这些内容将作为接下来系统实现的主要方向和依据，为系统实现提供了具体可行的方法。

# 影视票房预测系统实现

本章将根据上一章节详细设计的结果对系统进行开发，并最终展示系统的实现结果。本章将从影视评论分析、影视作品分析、系统整体呈现三个核心模块的实现情况，来进行整个系统的功能展示。

## 影视评论分析实现

### 影评情感分析实现

影评情感分析，综合表情符号、核心情感词典、分类器模型，最终在两级、三级情感分类上都取得了较高的准确率。具体判定过程，遵从以下算法流程：

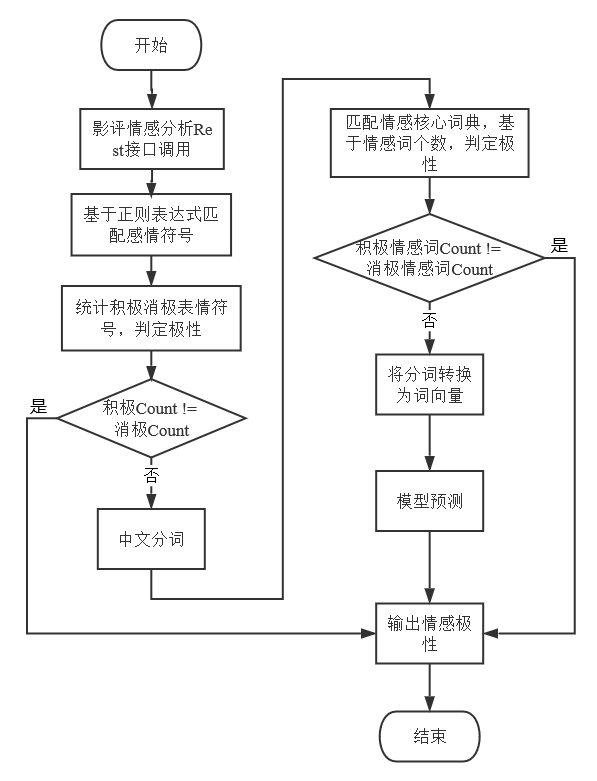
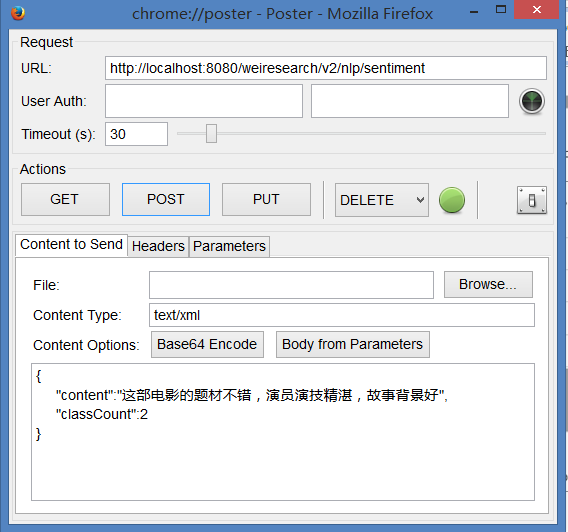
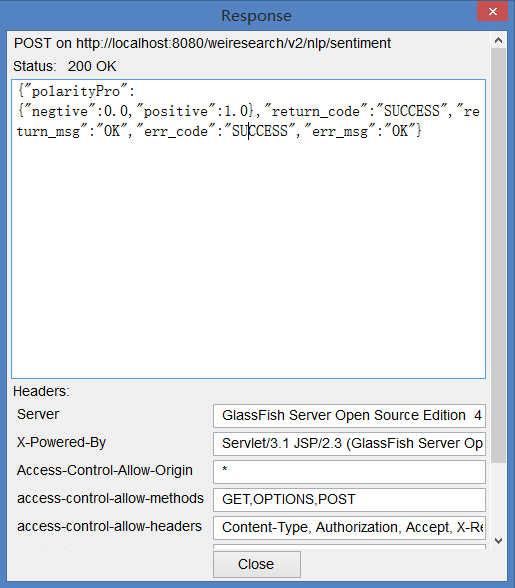


图4-1 影评情感分析算法流程

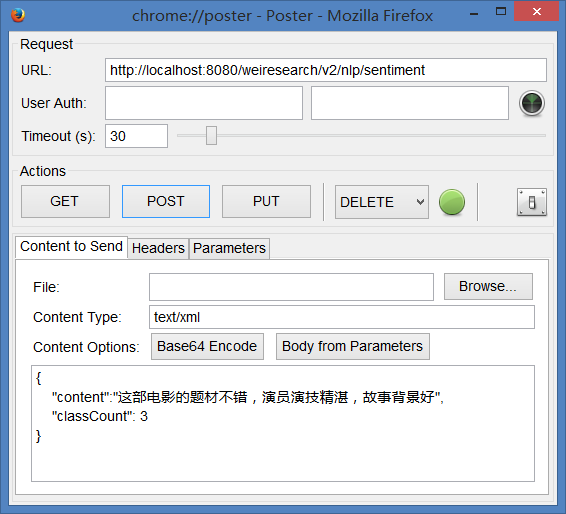
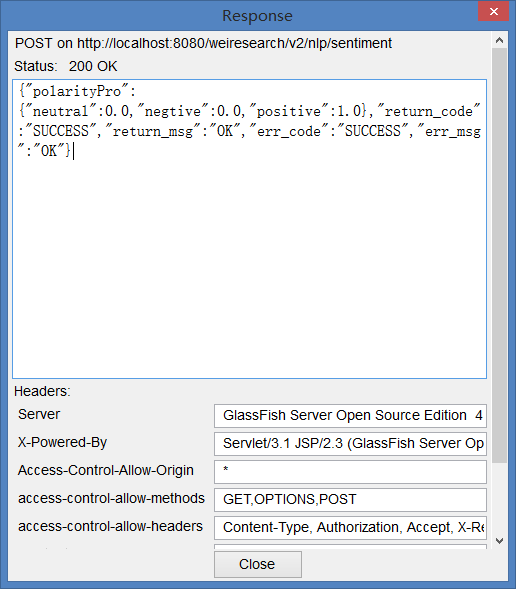
影评情感分析，目前实现了两级和三级情感划分两种分类模型，并封装成web service，以便调用。以文本“这部电影的题材不错，演员演技精湛，故事背景好”为例，积极、消极两级情感分类web service调用效果如图4-2所示：

（a）两级情感分类接口调用示例 （b）两级情感分类接口返回

图4-2 两级情感分类接口调用

基于同样的测试样例，积极、中性、消极三级情感分类web service调用效果如图4-3所示，相比二级分类接口，只需要调整classCount参数，默认情况下classCount为3，调用三级情感分类。

（a）三级情感分类接口调用示例 （b）三级情感分类接口返回

图4-3三级情感分类接口调用

整个影评情感分析具体集成到影视情报站项目中，呈现效果如图4-4所示，针对来自豆瓣、优酷、微博、时光网不同媒体平台上的影视评论，进行三级情感判定，并输出判定结果。



图4-4 影评情感分析集成效果

### 影评关键词提取实现

影评关键词提取在具体实现时，采用TextRank算法，具体实现过程，算法流程如下：

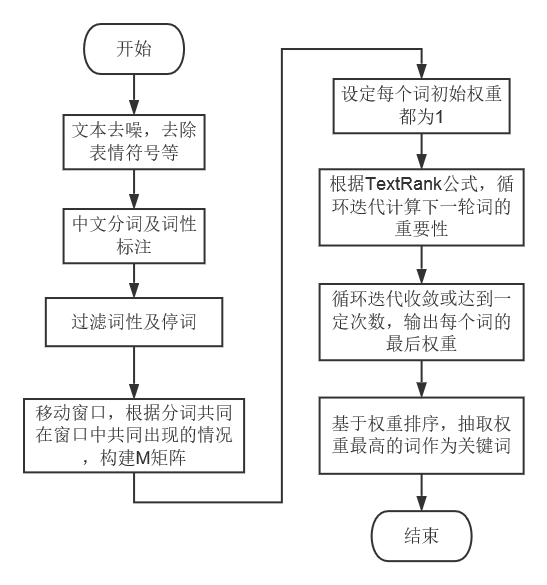
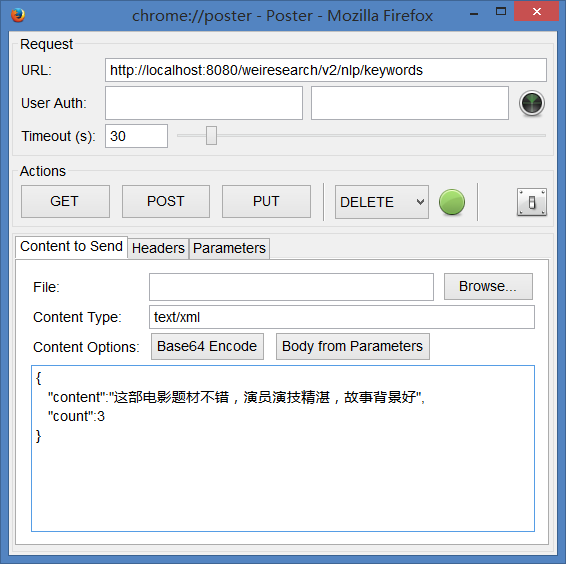
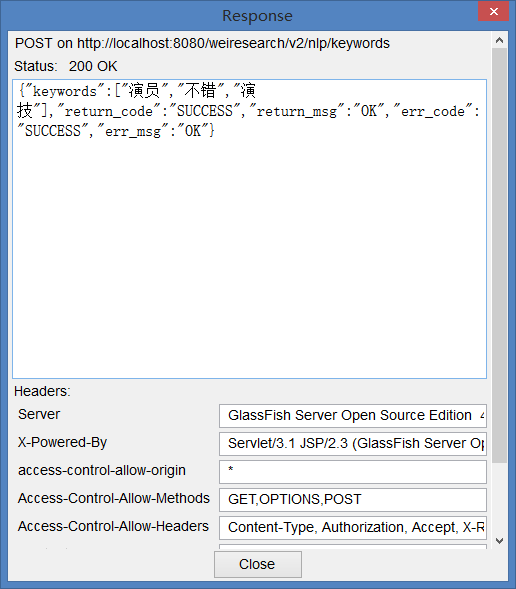


图4-5 TextRank实现算法流程

目前，影评关键词提取已经封装成web service接口，接口参数如下，需要给出待分析的文本content以及提取的关键词个数count，并封装成json格式进行提交，调用样例如图4-6所示：

（a）影评关键词提取调用示例 （b）影评关键词接口返回

图4-6 影评关键词提取接口调用

具体在集成到影视情报站项目中，会针对来自不同平台的影视评论，提取TOP 3关键词，呈现效果如图4-7所示：



图4-7 影评关键词集成效果

此外，抽取部分影评样本，基于影评的关键词，以及评论的情感极性，统计不同情感倾向影评中常出现的关键词，将出现频率较高的关键词可视化展现出来，展示效果如下：

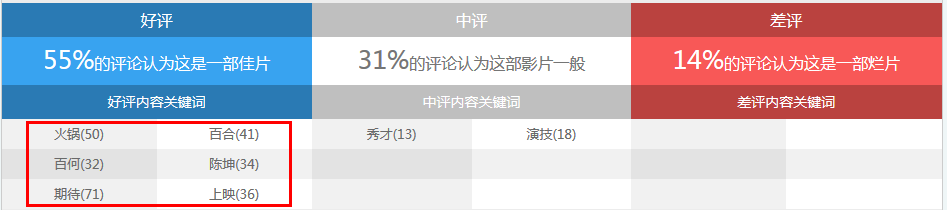


图4-8 不同情感倾向影评关键词

## 影视作品分析实现

### 影人主创指数实现

影人主创指数计算，主要基于公式(10)，具体在实现过程中，遵从以下算法流程：

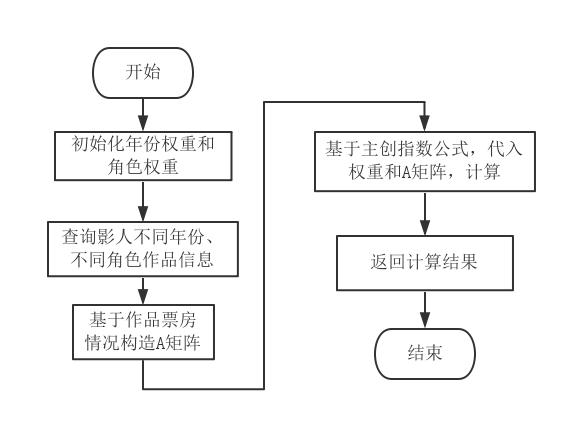
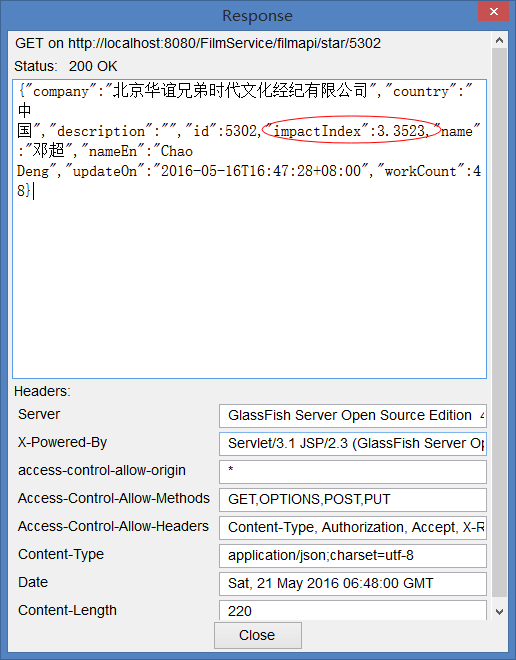


图4-9 主创指数计算算法流程

目前，针对影人的主创指数，基于jersey框架，进行了web service接口的封装，支持远程调用查询影人的基本信息和主创指数，查询样例如图4-10所示：

（a）影人信息查询调用 （b）影人信息接口返回

图4-10 影人接口调用示例

基于影人主创指数的计算结果，可以对主创人员进行主创排名，并基于影人身份以及地区进行筛选，进而分析当下最红最火的影人都有哪些，国内影人主创排名情况如图4-11所示：



图4-11 国内影人主创排名

### 影视作品受众人群分析实现

影视作品受众分析则从性别、收入、兴趣类别、学历、地域五个角度进行统计和分析。性别和地域信息基于网民的基本信息进行获取，收入、兴趣类别、学历信息则基于网民的基本信息和标签信息进行推算，虽然无法真实获取，但也具有一定可信性。

性别、收入、兴趣群体受众分析如图4-12所示：

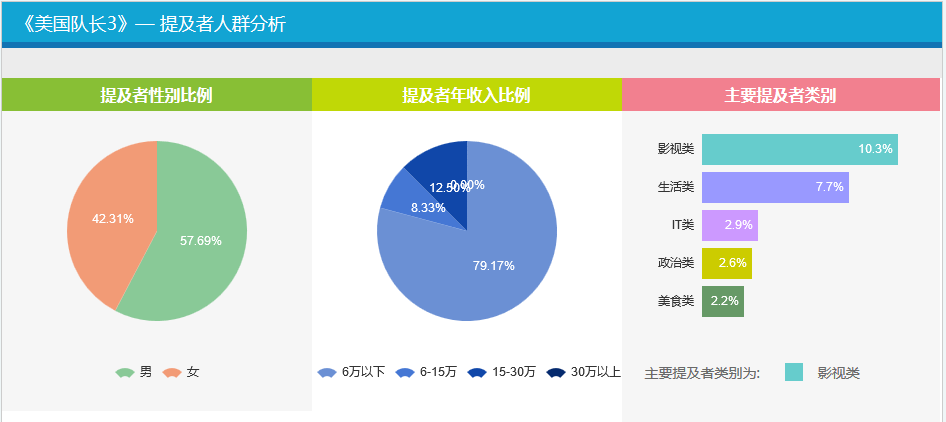


图4-12 影片受众人群分析

提及者学历受众分析如图4-13所示：

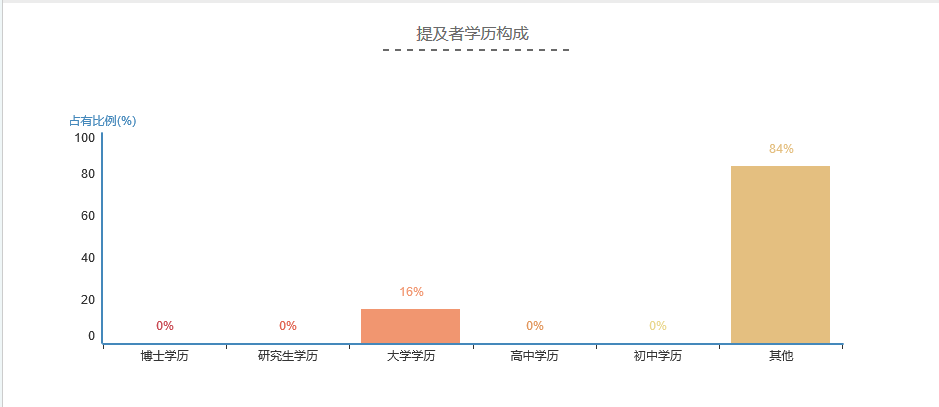


图4-13 影片受众人群学历分布

从国内一二线城市出发，基于用户信息，开展地域受众分析，分析结果如图4-14所示：

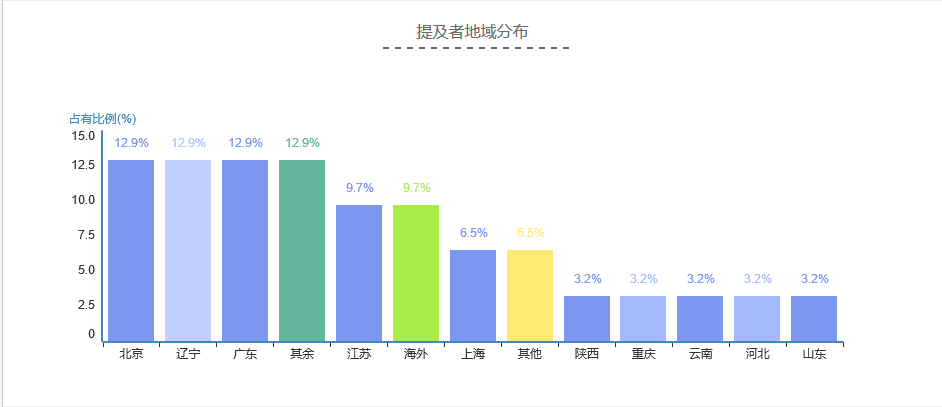


图4-14 影片受众人群地域分布

### 影视作品票房预测实现

在影视票房预测模型开展的过程中，考虑到观影人群消费能力的变化，以及影视票房的时间趋势，因此仅收集2011~2016年的影视作品信息，作为训练数据和验证数据，其中：

* 百万以下票房作品403部
* 百万票房作品456部
* 千万票房作品412部
* 亿万票房作品269部
* 十亿以上票房作品19部

在抽取影视作品数据时，从数据库中拿到的原始数据格式如图4-15所示：



图4-15 训练数据原始格式

每一行记录影片的基本信息、主创人员的角色信息及主创指数，基于以上数据格式，进一步进行整理和数值化，将影片类型、地区、票房等映射成数字，整理后的数据格式如图4-16所示：

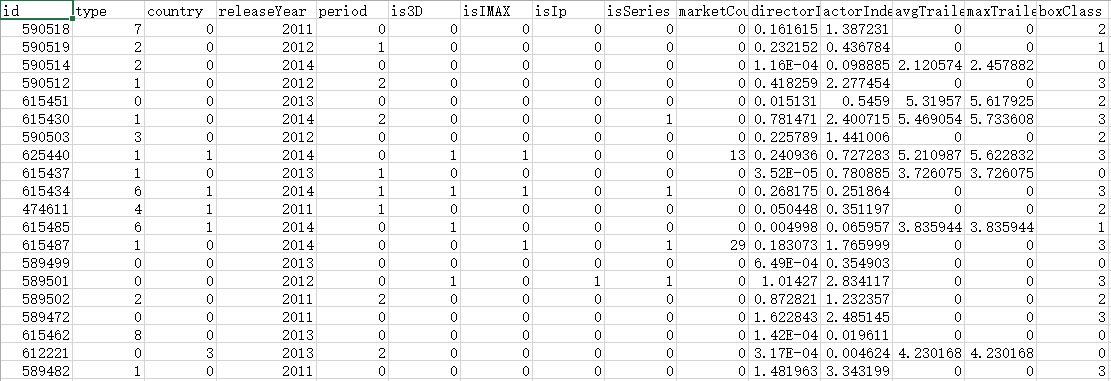


图4-16 训练数据数值化处理结果

基于整理后的数据，转换成weka可以直接使用的数据格式，以arff格式进行存储，每一部影片对应文件中一条训练数据，文件数据格式如图4-17所示：

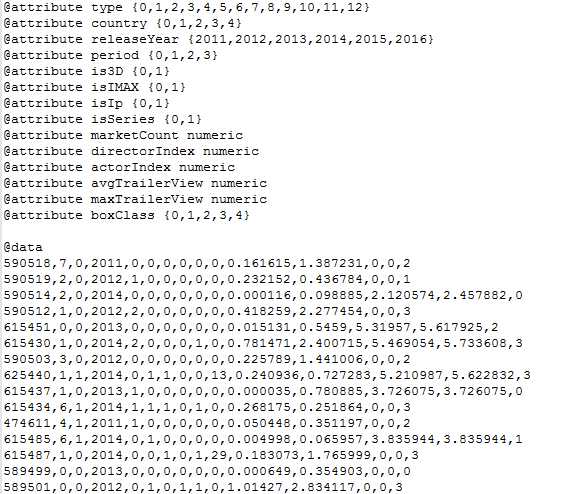


图4-17 影视票房数据arff文件

基于arff文件，尝试了支持向量机、朴素贝叶斯、决策树、随机森林、逻辑回归等分类模型。其中，J48决策树、随机森林效果、逻辑回归算法的性能和准确率最好，综合3种模型，进行模型投票，最终判定影片票房类别，最终可以达到70%左右的分类准确率。如图4-18所示，J48在验证集上的测试结果如下：



图4-18 决策树模型测试结果

基于上面提出的票房预测模型，进一步封装成web service，以“美国队长3”为例，基于影片名调用rest接口，调用接口和返回字段如图4-19和4-20所示：

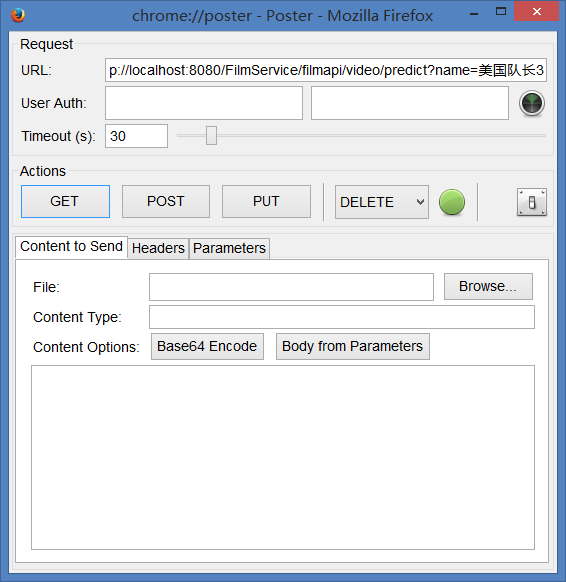


图4-19 票房预测Rest接口

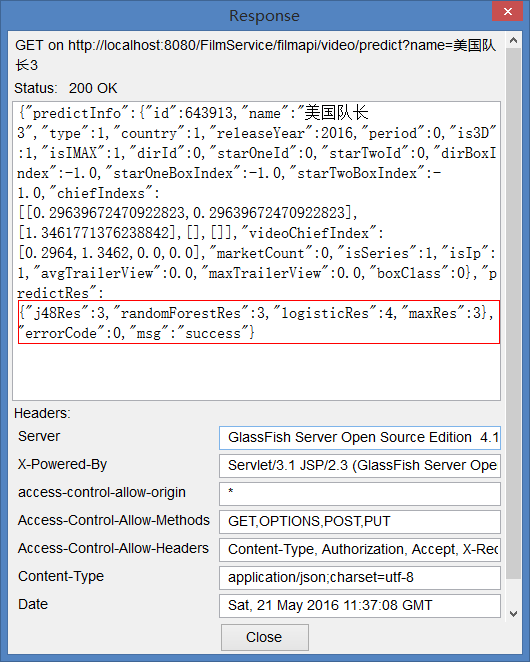


图4-20 票房预测Rest接口返回结果

预测接口返回JSON格式的字符串，其中predictRes字段记录不同分类器算法的预测结构信息，以“美国队长3”为例，该影片被J48分类器和随机森林分类器划分为亿万作品，逻辑回归则划分为十亿作品，最终根据投票结果，返回maxRes字段为亿万作品。

集成到影视情报站中，影视票房预测的展现效果则如图4-21所示：



图4-21 票房预测界面展现

## 系统整体可视化呈现

系统整体界面如图4-22~4-29所示，包括影视作品搜索、影视作品信息呈现、票房统计、排片情况、上座率情况、影片评论情况、受众人群分布等信息。

其中图4-22为影视作品的搜索界面，同时提取近期上映的5部影片，可以直接点击查看相关信息。



图4-22 影视作品搜索界面

图4-23为影片搜索结果界面，基于输入信息反馈相关影视作品及简要信息介绍。



图4-23 影视作品搜索结果

图4-24为影片基本信息情况，包括影片类型、上映时间、豆瓣评分、主创团队、剧情简介信息等。

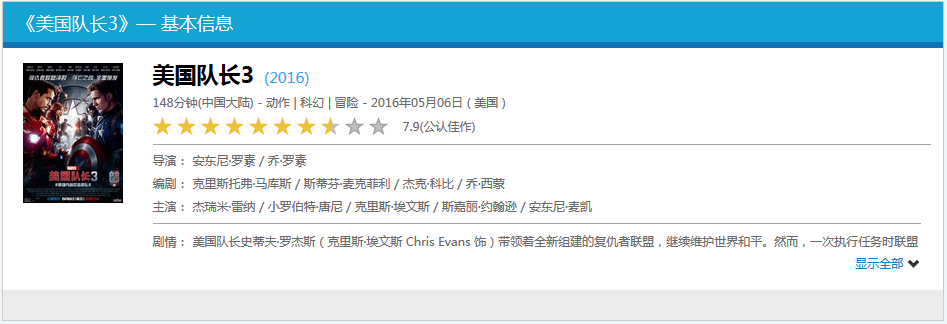


图4-24影视作品基本信息

图4-25为影片票房统计信息，信息来自猫眼电影和微票，包括影片上映首日的票房信息，当前周票房信息，以及截止到当前的票房情况。

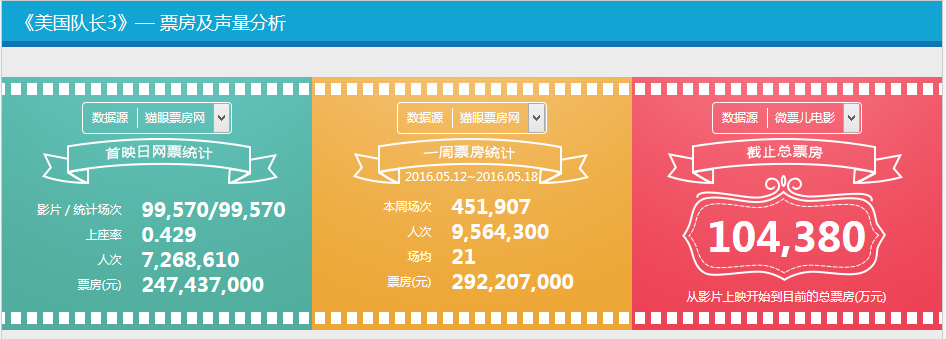


图4-25 影视票房统计

图4-26为影片的票房及声量，信息源来自猫眼电影、微票和微博，记录影片上映后每天的票房变化以及相应的微博声量。

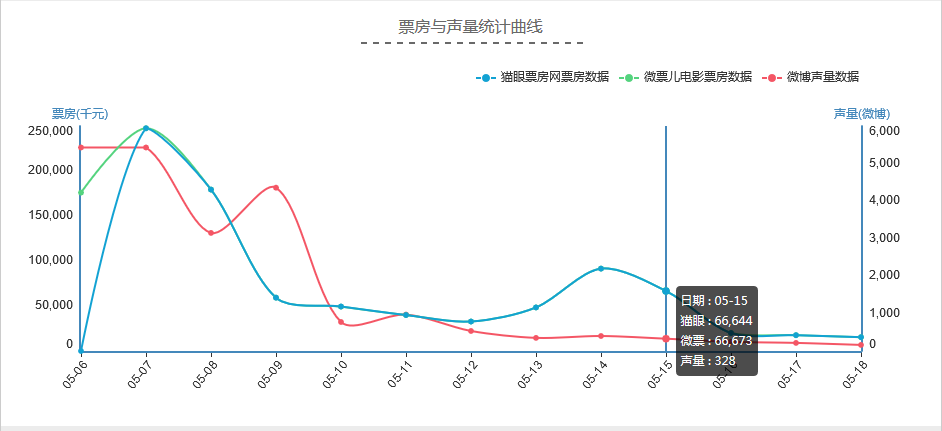


图4-26 影片票房与声量统计曲线

图4-27为影片的预告片信息，主要信息来源来自优酷，记录预告片的播放量、点赞量、点踩量信息。



图4-27 影片预告片信息

图4-28为影片的口碑情况，统计不同情感倾向评论的分布情况，以及分析影评关键词。



图4-28 影片口碑信息

图4-29为影片主创的社交账号信息，并对粉丝的性别及年龄构成进行了粉丝属性分析，自身热点话题也进行了统计。



图4-29 影人社交账号分析

## 本章小结

本章主要通过实现结果截图说明系统实现的结果，并通过算法实现流程图对重点功能的应用的算法的实现逻辑进行说明。至此系统的功能已完全实现，并为后续系统的测试做好了准备。

# 影视票房预测系统测试与性能分析

系统实现后仍需要进行测试和性能分析来确定系统是否能够正常完成相应功能。本章将叙述部分测试过程，对系统各主要功能进行系统测试，并给出相应的测试结果。此外，对系统的核心功能和算法，进行了性能分析，对比了不同平台及算法的性能，并给出相应的分析图表，进而验证系统现有算法和功能的有效性。

## 影视票房预测系统功能测试

### 影评情感分析功能测试

影评情感分析功能测试，则主要从数据的输入输出角度，来测试预测算法是否能够在给定待分析文本的情况下，正确判定情感极性，情感极性判定测试用例如表5-1所示：

表5-1 情感分析rest接口测试用例

|  |  |
| --- | --- |
| 功能模块： | 影评情感分析 |
| 用例说明： | 随机选取微博上的影视评论，指定情感分类，进行情感极性判定 |
| 测试用例： | 1. "content":"这部电影的题材不错，演员演技精湛，故事背景好","classCount":2 2. "content":"这部电影的题材不错，演员演技精湛，故事背景好","classCount":3 3. "content":"这部电影的题材不错，演员演技精湛，故事背景好","classCount":0 4. "content":"这部电影的剧情非常枯燥，演员演技也非常浮夸","classCount":3 |
| 基本事件流： | 1. 输入分析文本，及分类数 2. 正则表达式匹配“[]”中文本(微博表情)，基于微博表情判定极性 3. 若无法判定，基于核心情感词典进行判定 4. 若无法判定，基于classCount字段，调用两级或者三级分类器进行情感分类，若classCount !=2|3，默认采用三级情感分类 |
| 预期结果： | 返回不同情感的概率分布 |
| 实际结果： | 针对指定classCount为2|3，都能准确返回情感概率分布，针对classCount给定其他数值的情况，能够正确采用三级情感分类 |
| 结论： | 通过 |

### 影评关键词提取功能测试

影评关键词提取功能测试，则需要给出代分析文本，以及提取的关键词数，默认提取3个关键词，能够对中文以及英文文本实现关键词提取，测试用例设计如表5-2所示：

表5-2 关键词提取Rest接口测试用例

|  |  |
| --- | --- |
| 功能模块： | 影评关键词提取 |
| 用例说明： | 选取中文、英文文本，以及不同的关键词数，测试系统功能的正确性 |
| 测试用例： | 1. "content":"这部电影的题材不错，演员演技精湛，故事背景好","count":3 2. "content":"这部电影的题材不错，演员演技精湛，故事背景好","count":2 3. "content":"这部电影的题材不错，演员演技精湛，故事背景好","count":"" 4. "content": "This film is so interesting, it is so fun","count":3 |
| 基本事件流： | 1. 输入分析文本，及关键词数 2. 判定关键词的有效性，无效则默认采用3个关键词 3. 分词，基于分词序列，调用TextRank算法，计算词的重要性 4. 基于词的重要性，返回关键词 |
| 预期结果： | 根据关键词数，返回文本相应的关键词 |
| 实际结果： | 针对给定关键词的中英文文本，都能正确返回关键词，对于关键词个数为字符串格式的输入，能够调用默认关键词个数3进行分析 |
| 结论： | 通过 |

### 影人主创指数功能测试

影人主创指数功能测试，目前可以基于影人id和影人姓名进行查询，对于有效影人id和姓名，返回主创指数；对于无效id和姓名，返回数值-1，相应的测试用例设计如表5-3所示：

表5-3 主创指数接口测试用例

|  |  |
| --- | --- |
| 功能模块： | 影人主创指数 |
| 用例说明： | 选取有效、无效影人id和姓名，分别查询影人主创指数，查看相应返回结果 |
| 测试用例： | 1. 影人id：5302 2. 影人id：123456 3. 影人姓名：邓超 4. 影人姓名：123456 |
| 基本事件流： | 1. 基于给定id或姓名，验证影人的有效性 2. 无效，直接返回-1 3. 有效，查询或计算主创指数，并返回 |
| 预期结果： | 返回-1或者主创指数 |
| 实际结果： | 无效id和姓名，能够返回-1，有效id和姓名，能够正确返回主创指数 |
| 结论： | 通过 |

### 影片票房预测功能测试

影片票房预测，目前主要基于影片名称查询影片预测票房信息，对于有效的影片名，调用相应的分类器，开展票房预测，对于无效的影片名，返回字段进行提示，具体测试用例设计如表5-4所示：

表5-4 影片票房预测接口测试用例

|  |  |
| --- | --- |
| 功能模块： | 影片票房预测 |
| 用例说明： | 选取有效、无效的影片名，查询影片预测票房信息 |
| 测试用例： | 1. 影片名：“美国队长3” 2. 影片名：“123456” |
| 基本事件流： | 1. 基于输入影片名，查询影片是否有效 2. 无效，直接返回并提示 3. 有效，则查询相应信息，进行票房预测 |
| 预期结果： | 无效影片名，返回提示，有效影片名，能够返回票房预测结果，对于数据缺失的影片，也能够返回提示 |
| 实际结果： | 无效影片，在返回字段errCode和msg中会进行说明；有效影片，正确返回影片预测信息和票房预测结果 |
| 结论： | 通过 |

## 影视票房预测系统核心模块性能分析

### 影评情感分析性能分析

影评情感分析主要实现了两级和三级情感划分，具体在性能分析中，针对微博影评语料，对比了现有的腾讯文智平台及Bson平台，在算法的准确性和性能上都具备一定的优势，具体对比情况如图5-1所示：

图5-1 两级情感分析平台对比

由于腾讯文智和BsonNLP目前不支持中性判定的接口，但针对积极和消极的判定会给出概率分布，因此在这里，考虑设定概率阈值的方式，引入中性判定，判定规则为：

* 消极或者积极的概率 >= 0.6时，判定为极性
* 其余情形判定为中性

基于以上判定规则，测试结果如图5-2所示：

图5-2 三级情感分析平台对比

### 影视票房预测性能分析

在影视票房分类预测的过程中，对比了随着分类数的增加，不同学习算法性能的变化，在这里，我们抽取了分类数为2~5时，朴素贝叶斯、支持向量机、逻辑回归、决策树四种学习算法的性能变化，综合表现下来决策树的表现最佳，贝叶斯算法当分类数增加时，算法性能衰减最快。分类数为2~5时，对应的类标签说明如下：

* 分类数为2时：数据集划分亿万票房和非亿万票房两类；
* 分类数为3时，数据集划分亿万票房、千万票房、千万票房以下三类；
* 分类数为4时，数据集划分亿万票房、千万票房、百万票房、百万票房以下四类；
* 分类数为5时，数据集划分十亿票房、亿万票房、千万票房、百万票房、百万票房以下五类；

算法性能对比如图5-3所示：

图5-3 影视票房随分类数变化性能对比图

由上可知，当预测算法分类数仅为2时，算法性能最佳，可以达到90%左右的准确率，可以有效预测影视作品的票房是否成功，是否可以达到亿级大片的水平。而当分类数为5时，预测分类涵盖的票房种类则比较全面，更加具备实际应用意义，能够一定程度上预测影片的票房范围，辅助决策。

为了进一步研究分类数为5时的算法性能，我们进一步对比了J48决策树、随机森林效果、逻辑回归算法三类算法，相比其他算法，这三种模型的性能和准确率最为优越。在具体判定时，也考虑基于这三种模型，进行投票，进而判定影片票房类别。单一模型算法及投票算法性能对比如图5-4所示：

图5-4 影视票房五分类预测算法性能对比

整体来看，单一学习算法中，决策树具有较好的性能，但同时组合三种算法，基于投票的方式，最终输出影视票房预测结果，其算法的性能要更为优越，可以达到70.05%准确率，具有一定的提升效果，因此在具体应用时，则采用了组合不同学习算法，进行投票选择的方式。

## 本章小结

本章主要介绍系统测试工作，对系统中影评情感分析、影评关键词提取、主创指数、影视票房预测等各主要功能进行了功能测试，确认了功能的正确性。并对系统核心模块影评情感分析、影视票房预测进行相应的性能测试，验证了算法模型的性能和准确性。

# 结 论

经过系统的需求分析、设计到实际开发，最终完整实现了影视票房预测系统，并集成到影视情报站项目中。为了提高预测模型的准确率，提出了很多新的想法和思路，除了借鉴影视作品基本信息，还提出了影人主创指数的概念，并分析了社交媒体上的影视评论，判断影评的情感倾向，从而得出影片的口碑倾向。

在具体口碑分析实现的过程中，提出了综合表情符号、核心情感词典、分类器的情感判定新思路，能够有效实现影评积极、消极两级情感判定和积极、中性、消极三级情感判定模型。其中两级情感分类达到82.2%准确率，三级情感分类达到67.5%准确率，在影评情感分析领域，已经达到较高的准确率。

为了将影人信息数值化，更好地融入到影视票房预测中，提出了主创指数的概念，能够有效基于影人过去几年内以不同身份参与影片取得票房的情况，分析影人的票房影响力。并且在对主创指数排序的过程中，也可以有趣的发现国内外目前的影人排名，并且排名具有较好的准确性和一致性。

影片票房预测过程中，则综合了影片基本信息、影人主创指数、影片口碑信息，提出了基于决策树、随机森林、逻辑回归的投票式票房预测模型，在对2016年影片开展分类预测过程中，两类预测可以达到90%准确率，五类预测可以达到70%左右的准确率，已经具备较高的实用价值。

受限于数据的因素，整个模型还是有些许不足的地方，在实现过程中，也考虑到引入影片预告片的播放量信息、营销事件信息、社交平台想看量信息、百度搜索指数信息等，但很多数据信息很难追溯以及较难获取，导致整个模型还有很大的提升空间。所以，在今后继续完善的过程中，可以从这些数据维度入手，挖掘新的数据维度对影视票房的影响，提高模型预测的准确度。

最后，整个项目集成到了公司现有的影视情报站项目中，使影视情报站能够完整地展现影片的各个方面的信息，涵盖基本信息、排片及上座率信息、预测票房信息、口碑信息、受众人群信息、粉丝属性信息，从而为用户提供全方位多角度的影视服务。

# 参考文献

[1] Apala K R, Jose M, Motnam S, et al. Prediction of movies box office performance using social media[C]//Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on. IEEE, 2013: 1209-1214.

[2] Du J, Xu H, Huang X. Box office prediction based on microblog[J]. Expert Systems with Applications, 2014, 41(4): 1680-1689.

[3] 史伟, 王洪伟, 何绍义. 基于微博情感分析的电影票房预测研究[J]. 华中师范大学学报：自然科学版, 2015, 49(1):66-72.

[4] 胡小莉, 李波, 吴正鹏. 电影票房的影响因素分析[J]. 中国传媒大学学报：自然科学版, 2013, 第1期(01):62-67.

[5] Thelwall M, Buckley K, Paltoglou G, et al. Sentiment strength detection in short informal text[J]. Journal of the American Society for Information Science and Technology, 2010, 61(12): 2544-2558.

[6] Wang J, Song D, Liao L, et al. The Chinese Bag-of-Opinions Method for Hot-Topic-Oriented Sentiment Analysis on Weibo[M]//Semantic Web and Web Science. Springer New York, 2013: 357-367.

[7] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques[C]//Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002: 79-86.

[8] 刘鲁, 刘志明. 基于机器学习的中文微博情感分类实证研究[J]. 计算机工程与应用, 2012, 48(1):1-4.

[9] Wilson T, Wiebe J, Hwa R. Just how mad are you? Finding strong and weak opinion clauses[C]//aaai. 2004, 4: 761-769.

[10] 董喜双, 关毅, 李本阳, 等. 基于最大熵模型的中文词与句情感分析研究[C]. 第二届中文倾向性分析评测. 上海, 2009: 50-58.

[11] Zhai Z, Xu H, Kang B, et al. Exploiting effective features for chinese sentiment classification[J]. Expert Systems with Applications, 2011, 38(8): 9139-9146.

[12] Nair D S, Jayan J P, Rajeev R R, et al. Sentiment Analysis of Malayalam film review using machine learning techniques[C]//Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on. IEEE, 2015: 2381-2384.

[13] Kim D, Kim D, Hwang E, et al. A user opinion and metadata mining scheme for predicting box office performance of movies in the social network environment[J]. New Review of Hypermedia and Multimedia, 2013, 19(3-4): 259-272.

[14] Huang S, Peng W, Li J, et al. Sentiment and topic analysis on social media: a multi-task multi-label classification approach[C]//Proceedings of the 5th annual ACM web science conference. ACM, 2013: 172-181.

# 哈尔滨工业大学本科毕业设计（论文）原创性声明

本人郑重声明：在哈尔滨工业大学攻读学士学位期间，所提交的毕业设计（论文）《基于情感分析与主创指数的影视票房预测系统设计与实现》，是本人在导师指导下独立进行研究工作所取得的成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明，其它未注明部分不包含他人已发表或撰写过的研究成果，不存在购买、由他人代写、剽窃和伪造数据等作假行为。

本人愿为此声明承担法律责任。

作者签名： 日期： 年 月 日

# 致 谢

衷心感谢导师涂志莹老师对本人的精心指导。在毕业设计进行的过程中，遇到过很多困难和疑惑，涂老师都给予了我无私的指导和帮助，尤其记得在开题的时候，涂老师基于我当时所作的工作为我拟了很多的题目，并给出后续可做的方向，从而能够让我清晰立题和明确将来所作的方向。在后续过程中，涂老师也多次细心为我批改论文和指出改进意见。他的言传身教使我终身受益。

感谢在实习公司微瑞思创的校外导师以及各位同事，从你们身上学到了很多专业技能和为人处世的道理，多亏了你们，在我毕设遇到困难时，给予我支持和发散的思路，让我能够克服实现过程中的种种困难。

感谢这篇论文所涉及到的各位学者。正因为各位前人的经验和成果，使我萌生了许多思路和想法，也更加明确了研究的方向。

感谢我的同学和朋友，在我遇到困难时，给予我鼓励和支持，陪伴我度过了许多难关。

感谢家人对我的理解与支持，使我能够专心完成毕业设计与毕业论文的相关工作。

最后，谢谢大学四年一起陪伴过来的所有人，谢谢大家。

# 附录1 外文文献译文

基于社交平台的影视票房预测

Krushikanth R. Apala, Merin Jose, Supreme Motnam, C.-C. Chan, Kathy J. Liszka, and Federico de Gregorio

Department of Computer Science

Department of Marketing1

The University of Akron

摘要：

在此次研究中，我们使用来自推特、YouTube、IMDb等不同社交媒体及网络平台的影视数据，采用数据挖掘的方法，在影视票房预测上挖掘到了一些有趣的知识。整个影视票房预测基于影视历史数据库信息、推特上的关注量信息、YouTube上的影评情感分析信息等关键因素。基于K-means聚类的方式将影视作品进行标注，并且划分为成功、中等、失败三个等级。模型训练时，则采用了weka中的J48工具，并发现了影视票房预测中相应的规则。整个票房预测针对2013年夏季未上映的影片，具体的研究结果和性能请参考本篇论文中后续的研究内容。

关键词：情感分析；数据挖掘；YouTube；推特；IMDb；预告片

1 引言

社交媒体平台Twitter, YouTube等，一直以来都是人们在日常生活中针对各个事件分享内容和讨论意见的平台。显而易见，如果能把这些信息整合进来，辅助整个决策过程，将会带来很大的商业收益。例如，基于社交评论衍生的影视票房预测模型就能够为影视作品制作人员的决策带来很大帮助。

在社交媒体的用户社区中，影视作品的主题具有相当可观的利益。目前，很多研究用来提出影视收入的预测模型。在文献1中，基于Twitter的线性回归模型已经被用来预测影视票房。其中，推特常常作为研究影视票房收入总量的炒作因素。也有一些研究，基于维基百科的信息来预测票房表现。IMDb也是这些类型研究中另一大非常火的数据源。在论文15中，影视作品的预测基于影评中的一些关键点。同样在论文16到19中，影视预测基于IMDb的数据，同时采用了回归模型或者概率模型。在论文20中，一种基于网络数据挖掘的方法被引入来预测2006年影视作品是否成功以及能够冲击奥斯卡奖项。该方法结合了社交平台分析以及自动的评论情感分析。并且在影视行业中的趋势以及真实事件成功被预测基于IMDb上的论坛发帖。参考论文21，描述了一种基于IMDb上240部影片，并且采用数据挖掘的方式，借助weka中的J48算法提出了一种影片打分的原始模型。在论文22中，对预测影片受欢迎程度的方法模型进行了讨论，并且衍生出指导内容分发网络的一种新策略。演员和导演的人气被作为影片受欢迎度预测的最基本的影响因素。数据挖掘的方法发现影片质量与影片投入预算的相关性，同时，也有相关研究发现了其他关于成功影片至关重要的因素。另一项非常有趣的工作是应用图像处理和自然语言处理，并将其转换成可计算的系统。提出的系统通过考虑影片中的动作与言语等因素来实现影片打分。但是系统的准确率只有66.67的水平。影视评论的两级分类在论文24中有所讨论，影评数据来自两大数据集，并且被划分为积极、消极两类。YouTube是最受欢迎的社交平台之一。上百万的用户每天在平台上上传、分享、讨论影片的相关内容。在YouTube平台近期的研究中，每分钟大约有72小时长的影片内容会被上传。

在这项工作中，我们生成一个预测模型，考虑了IMDB电影库中用户收视率排名较高的电影类型，导演的人气指数，领衔主演的Twitter的粉丝量，借助YouTube上官方预告片观看量和评论量反映的影片人气情况，以及从评论中反馈的情感倾向等因素。我们还考虑影片是否是序列电影。总共，我们从社会化媒体数据提出了8个数据维度来对评估未上映的影视作品。所有的属性值都被映射成一个整数，通过简单的最小-最大归一化处理的方法，将数据维度映射成1~10中的数值。然后，Weka中的K均值聚类工具[14]被用于生成电影的类别，聚类为三类，具体标签为：成功，中等或失败。 最后，我们使用Weka中的J48决策树分类器训练得到预测模型。

在我们的实验中，我们收集了2013年5~7月中上映的35部影片在2月3月的社交媒体的数据。有趣的是，在J48分类器中，八个属性中只有四个被认为影视作品是否成功有关，即男女主演的人气情况、影片的类别和续集情况。在今天夏天影片真正上映的时候，最终结果才能够被验证，并且到时我们会更新相应的结果。

在下面的部分，我们将描述具体数据的采集过程。第三节介绍实验开展的情况。结果与讨论会在第四节中提出，并随之提出我们的结论。

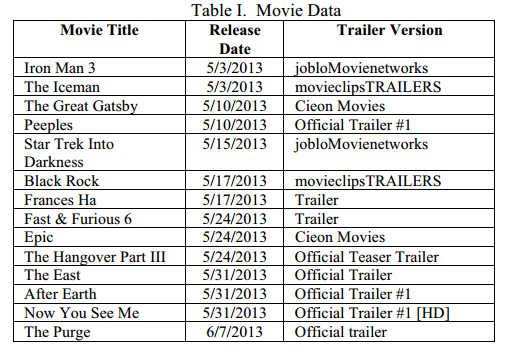
2 数据采集

我们收集了2013年5~6月美国市场上映的影片官方预告片数据。我们基于网站http://www.comingsoon.net[7] 确定了这批影片的列表。明确影片后，我们从http://themoviebox.net[8]网站跟踪了这批影片的电影预告片。从预告片的数据中，我们可以确定这些电影预告片在YouTube上的视频ID号。 整个35部影片预告片名和发行日期如表1所示：

对于YouTube上的开发者指南是提供不同语言的支持，如Java，.NET，PHP，Python和JavaScript。对于YouTube上API进行交互，我们基于PHP语言书写了相应的代码，并且运行在WAMP服务器上。所有在YouTube视频都有相应的视频ID。我们通过传递视频ID给我们的代码，便能够获取相应的视频评论。一旦执行代码，它使用YouTube提供的API接口提取视频评论，并且以文本文件的格式存储到本地。我们使用开源Zend框架，基于PHP语言开发了网络应用程序和相应的web服务。 其中，Zend框架完全面向对象，并且利用了很多PHP语言的新特性，如命名空间，延迟静态绑定，lambda函数和闭包[9]等。

YouTube的数据采集从2013年2月28日至3月28日进行了为期一个月左右时间。我们了提取影片每天观看次数和评论量。由于YouTube API的限制，我们一次最多只能获取最近时间内的1000条评论[9]。最终，关于35部选取的影片，我们采集到的数据，总播放量超过7,7000,000以上，总评论量98K。

在推特上，影片导演们的粉丝量接近137K，男演员的粉丝量达到7600,000，女演员达到20,800,000。

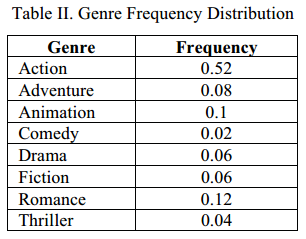




接下来，我们继续描述数据的属性特征时如何被提取的。

1. 影片类型

IMDb影片数据库是最全面的数据资源库，几乎包含每一部制作影片的详细信息。它包含海量的数据，能够从中获取关于影视总体趋势非常有价值的信息[10]。电影被划分为不同类型，如动作，冒险，动画，喜剧，戏剧，科幻，爱情和惊悚片等。我们选取了2010、2011和2012年TOP50打分的影视作品。表二给出了50部电影类型题材分布的情况，该数据被用于评估影视作品类型的受欢迎程度。



1. 导演、男女主演受欢迎度

导演，男女主角的受欢迎程度通过Twitter账号的粉丝量进行评估[12]。在我们收集YouTube数据同时，收集粉丝量信息。表三中展示了我们选取的35部影片的导演及男女主演的粉丝量情况。值得注意，在我们的数据集中，各有一个男演员和女演员没有Twitter账号，因此相应的数值设置为零。这可能会为相应的电影预测产生相应的误差。

1. 观看量和评论量

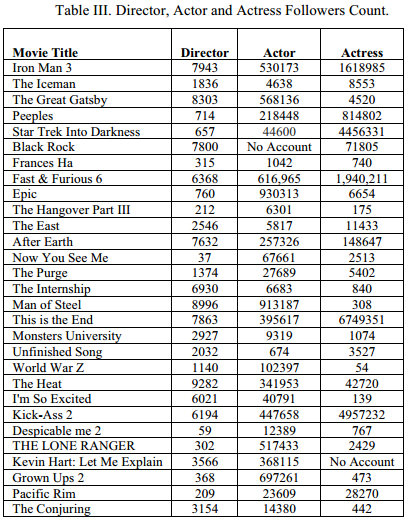
我们借助未上映影片官方预告片的播放量和评论量来评价影片的受欢迎程度。在同一时间，我们收集了YouTube平台上的影片播放量和评论量数据。

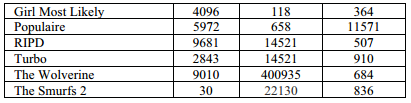
1. 系列电影

如果一部影片时某一部已上映影片的后续作品，那面该影片便被定义为系列电影。该数据维度只需要二值属性即可，“是”或者“否”。序列电影列表可以从IMDb中获取[10]。当计算决策权重时，一部系列电影会比一部从未上映过的影片更受观众欢迎。我们分配权值10给系列电影，而对于非系列电影，非配权值为1。

1. 情感分析

情感分析被用于YouTube整个评论数据集。我们参考论文[13]中收集能够标识情感的单词列表。数据集被处理成使用一组积极和消极词标识的形势，然后影视评论呗划分为积极，消极以及中性。整个处理过程遵从以下流程。首先，真个数据集被分成一群单个的影评。然后，每条评论通过分词器分词。分词后，每个分词与核心情感词词典进行比较。情感标定基于评论中匹配的不同情感词出现的频率进行判定。中性情感则意味着匹配的积极与消极情感词几乎相同。





3 实验过程

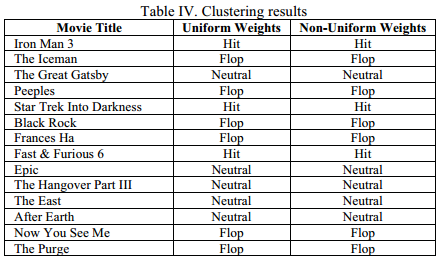
在整个试验中，主要有三个流程：归一化训练数据，应用K均值聚类生成类标签，训练预测模型。

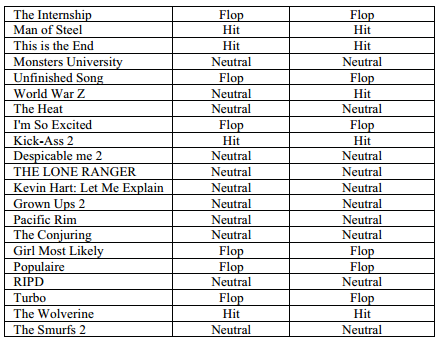
对于归一化数据，我们采用了简单的MIN-MAX[11]方法，接下来，采用四舍五入的方法将每个数据属性映射成1~10的整数值。

其中，是映射前的原始数据，是归一化的值。和是原始值的真实最大值和最小值。归一化处理后数据的最大值和最小值是和。

对于数据聚类，我们使用weka中K-means方法。所有影片数据被应映射成3类：成功，中等和失败。在聚类之前，我们考虑两种分配权重给属性的方式。一种方法是考虑所有属性，并且分配相同的权重，另一种是给予较高的权重给情感属性，其余属性权重相等。聚类结果如表4所示，其中我们看到，只有一部电影“世界大战Z”，在不同的数据权重分配方式下，标签不同。借助均权重的加权方法，相应标签为中性，在提高情感权重的方法下，相应标签为成功。通过进一步探索，这部电影预告片确实有大量的播放量及评论数，并且有大量积极评论，尽管它的女主角的粉丝量较低。

聚类结果被用于标记每部电影的预测值来创建训练集，进而提出后续的预测模型。我们使用Weka中的J48和朴素贝叶斯算法，建立了3折交叉验证验证的两种模型。在下面的章节介绍相应的实验结果。



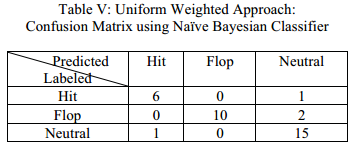


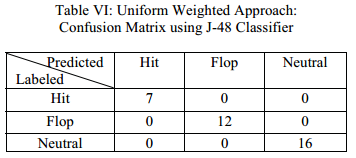
4 实验结果

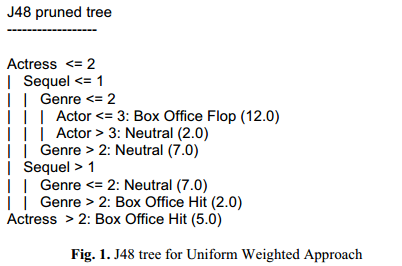
1. 均权重方法

在这种方法中所有属性被赋予相等的权重。35部电影中，7部电影被聚类为一类，标记为成功；12部电影被聚为一类，标记为失败；16部电影聚类为一类，标记为中等，这意味着预测结果也并不是确定的。聚类标签是基于属性值总分进行判定，8个数据位维度，每个维度满分10分，总分80分。每一类影片的打分临界值情况如下：(1)成功：在42至70的范围内，(2)中等：在21至33的范围内，及(3)失败：在9至20内。

然后，基于标记数据集，采用Weka中朴素贝叶斯和J48分类器生成预测模型并进行测试。所得混淆矩阵如表V和VI所示。通过J48产生的决策树如图1所示。从图1中我们可以看到，目前的决策树中的关键属性是女主角的受欢迎度，影评情感倾向，影片类别和导演影响力。我们还可以观察到，女演员的受欢迎度是J48决策树进行分类最相关的根节点因子。



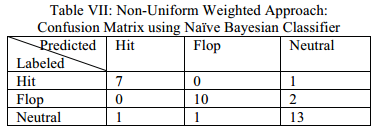


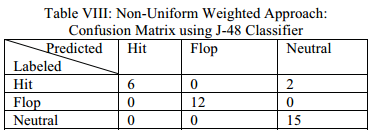


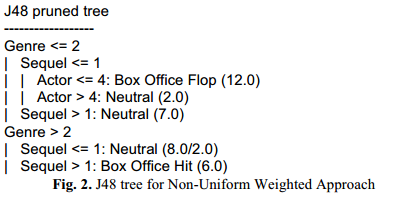
1. 非均权重方法

在这种方法中，更多的权重被赋给情感倾向数据维度。35部电影中，8部影片聚类在一起，标记为成功；12部影片聚类到一起，标记为失败；15影片聚类到一起，标记为中等。类标签的赋值基于属性维度打分，总分100分，30分为情感属性，1其余七个属性每个10分。每个类标签分类临界值如下：(1)成功：在42至66分的范围内，(2)中等：在22至35分的范围内，(3)失败：在13至19分氛围内。

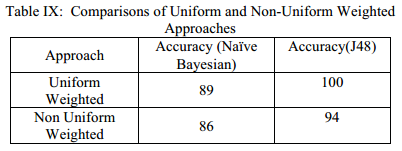
基于Weka分类和测试结果如表VII和VIII所示。使用J48生成的决策树如图2所示。从图2中，我们可以观察到相应的影响因素有影视类型，影视是否续集和男主演的受欢迎程度。我们也可以看到，影片类型在J48决策树中被认为是最相关的决策属性。

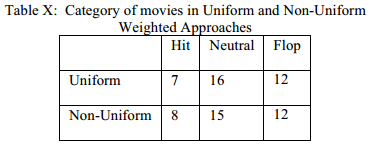






基于weka均权重和非均权重方法所得到误差统计结果如表IX所示。相应标签聚类结果如表X所示。





1. 初步结果

在我们的名单上的35电影，其中15部在2013年6月7日上映。我们从IMDb Pro电影网站http://pro.imdb.com和http://www.boxofficemojo.com上跟踪影片的票房情况。两部上映的影片“Black Rock”和“Frances Ha”，并没有被列在最初的35部影片中；因此，被排除在我们的初步结果中。表XI列出了网站上标有票房收入的电影名。对于每部电影中，我们跟踪电影的投资情况及每周全球票房和美国周票房。从预算和总收入水平，计算净利润和利润的比率。表XII列出了表XI中提及影片的初步结果。列WK给出了截止到2013年6月7日影片上映的周数。列Pred.表示模型预测的结果，其中N表示中等。WP列表示全球发行所获得的利润，而WPR是全球发行利润率。同理，USP是影片在美国上映获得的利润，USPR是在美国上映获得的利润率。利润显示单位为百万美元，括号表示负利润。

表XII显示只有一部电影“The Purge”被我们预测错误。在我们分类为中性的五部电影中，其中一部似乎趋于成功，两部趋向于失败，还有两部电影在全球票房和美国票房上有着混合利润。

Table XI: Partial list of movie titles

|  |  |
| --- | --- |
| Movie | Title |
| 1 | Iron Man 3 |
| 2 | The Iceman |
| 3 | The Great Gatsby |
| 4 | Peoples |
| 5 | Star Trek Into Darkness |
| 6 | Fast & Furious 6 |
| 7 | Epic |
| 8 | The Hangover Part III |
| 9 | The East |
| 10 | After Earth |
| 11 | Now You See Me |
| 12 | The Purge |
| 13 | The Internship |



5 结论

在本文中，电影的票房收入预测模型借助于来自社交媒体和IMDb数据源，从中挖掘中影响作品收入关键因素，用于预测影视作品是否成功。根据我们的模型，我们已经确定了以下的模式：（1）女主角的受欢迎程度是一部电影成功的关键;（2）结合过去卖座的影视类型和续集电影是票房成功的另一大关键；（3）一部知名度并不高的新片以及演员阵容的人气也一般，很有可能导致电影票房的失败。令人惊讶的是影评情感情况以及预告片的播放量和评论数与我们最终提出的模型相关度不高。我们认为，数据维度权重的分配与票房收入预测具有一定的相关性。因此，在后续的研究中，会探讨哪种权重分配方式对于最终的结果将收益最大。此外，我们的预测针对尚未公布的电影。目前跟踪13部电影的初步结果已经表明，我们的模型具有很好的预测性能。当所有影片都上映时，我们会最终验证我们的模型并且公布出来。为了改善我们的模型，今后开展的工作主要有将基于票房收益及收益率更加细化我们的分类模型，提升算法模型的性能。

# 附录2 外文文献原文







