原型图设计说明文档

本文档对原型图进行进一步补充说明。

# 摘要

近年来，随着电影行业迅猛发展，年度电影总票房持续快速增长。对电影票房进行预测可以给电影的制片方和宣发方提供票房数据支持，并在此基础上制更加完善的营销和宣传策方案。传统的电影票房预测往往由各大院线的相关工作人员完成。这种由人工完成的票房预测大多存在预测精度低、预测主观性强、人力资源成本高等问题，因此不能较好地满足实际需求。

针对人工进行票房预测存在的种种问题，本文设计了一个智能化的电影票房预测系统。该系统使用Python爬虫技术对外网电影相关数据进行采集，并对这些数据进行解析、过滤、转换等预处理操作，然后使用HDFS（Hadoop Distributed File System）文件系统结合Mysql数据库对完成预处理操作后的数据进行存储。系统的特征抽取模块利用特征工程相关技术和算法对每部电影的特征进行量化计算和数值修正。最后，系统使用普通最小二乘法、局部加权线性回归算法、随机森林算法和随机梯度回归树四种机器学习算法进行单模型学习训练，并对每个算法最好的模型进行算法融合，得到最终电影票房预测结果。

本文的最后使用2013年1月到2016年4月总共947部电影进行算法模型的训练与评测，并以判定系数R2值作为评判标准。系统最终的R2值达到0.9301，表明系统具有较高准确率。

**关键字：**电影票房，预测，大数据，Scikit-learn，机器学习

**目录**

[摘要 I](#_Toc459326582)

[Abstract II](#_Toc459326583)

[目录 III](#_Toc459326584)

[第1章绪论 1](#_Toc459326585)

[1.1 课题的背景与意义 1](#_Toc459326586)

[1.1.1 选题的依据 1](#_Toc459326587)

[1.1.2 选题的意义 2](#_Toc459326588)

[1.2 国内外文献资料综述 3](#_Toc459326589)

[1.2.1 电影票房预测研究进展 3](#_Toc459326590)

[1.2.2 大数据处理系统调研 4](#_Toc459326591)

[1.2.3 机器学习算法库调研 5](#_Toc459326592)

[1.3 本人的主要工作 6](#_Toc459326593)

[1.4 本文的组织结构 7](#_Toc459326594)

[第2章相关技术与理论分析 8](#_Toc459326595)

[2.1 Scikit-learn机器学习算法库 8](#_Toc459326596)

[2.2 Apache Hadoop 9](#_Toc459326597)

[2.2.1 Hadoop HDFS 9](#_Toc459326598)

[2.2.2 MapReduce计算框架 10](#_Toc459326599)

[2.3 Python爬虫技术 11](#_Toc459326600)

[2.4 Beautiful Soup 12](#_Toc459326601)

[2.5 特征选择技术 12](#_Toc459326602)

[2.6 预测算法 14](#_Toc459326603)

[2.7 本章小结 17](#_Toc459326604)

[第3章需求分析 18](#_Toc459326605)

[3.1 功能性需求 18](#_Toc459326606)

[3.1.1 需求概述 18](#_Toc459326607)

[3.1.2 需求建模 18](#_Toc459326608)

[3.2 非功能性需求 22](#_Toc459326609)

[3.2.1 准确性需求 22](#_Toc459326610)

[3.2.2 同步性需求 23](#_Toc459326611)

[3.2.3 健壮性需求 23](#_Toc459326612)

[3.2.4 可扩展性需求 23](#_Toc459326613)

[3.2.5 其它需求 23](#_Toc459326614)

[3.3 可能性分析 23](#_Toc459326615)

[3.4 本章小结 24](#_Toc459326616)

[第4章概要设计 25](#_Toc459326617)

[4.1 系统整体架构 25](#_Toc459326618)

[4.2 系统功能模块划分 27](#_Toc459326619)

[4.3 本章小结 31](#_Toc459326620)

[第5章详细设计与实现 32](#_Toc459326621)

[5.1 用户登录模块详细设计与实现 32](#_Toc459326622)

[5.1.1 用户登录整体流程 32](#_Toc459326623)

[5.1.2 静态类设计和核心接口实现 33](#_Toc459326625)

[5.2 电影数据查询模块 35](#_Toc459326626)

[5.2.1 静态类设计和核心接口实现 35](#_Toc459326627)

[5.3 数据抓取模块 36](#_Toc459326628)

[5.4 数据预处理模块 38](#_Toc459326629)

[5.4.1 数据解析 38](#_Toc459326630)

[5.4.2 数据存储 39](#_Toc459326631)

[5.4.3 数据过滤 39](#_Toc459326632)

[5.5 特征抽取模块 39](#_Toc459326633)

[5.5.1 离线特征计算 40](#_Toc459326634)

[5.5.2 特征数值修正 44](#_Toc459326635)

[5.6 算法预测模块 45](#_Toc459326636)

[5.6.1 数据集划分 46](#_Toc459326637)

[5.6.2 普通最小二乘法 46](#_Toc459326638)

[5.6.3 局部加权线性回归 46](#_Toc459326639)

[5.6.4 随机梯度回归树 47](#_Toc459326640)

[5.6.5 随机森林 48](#_Toc459326641)

[5.6.6 算法融合 48](#_Toc459326642)

[5.7 本章小结 50](#_Toc459326643)

[第6章系统部署与测试 51](#_Toc459326644)

[6.1 测试环境 51](#_Toc459326645)

[6.1.1 硬件环境 51](#_Toc459326646)

[6.1.2 软件环境 51](#_Toc459326647)

[6.2 系统部署安装 52](#_Toc459326648)

[6.3 功能测试 54](#_Toc459326649)

[6.3.1 用户登录功能测试 54](#_Toc459326650)

[6.3.2 电影数据查询功能测试 54](#_Toc459326651)

[6.3.3 数据抓取功能测试 56](#_Toc459326652)

[6.3.4 数据预处理功能测试 57](#_Toc459326653)

[6.3.5 特征抽取功能测试 57](#_Toc459326654)

[6.3.6 算法预测结果评估 60](#_Toc459326655)

[6.4 性能测试 63](#_Toc459326656)

[6.5 本章小结 64](#_Toc459326657)

[第7章总结与展望 65](#_Toc459326659)

[7.1 论文工作总结 65](#_Toc459326661)

[7.2 不足与展望 66](#_Toc459326662)

[参考文献 67](#_Toc459326663)

[致谢 69](#_Toc459326664)

# 第1章绪论

## 1.1 课题的背景与意义

### 1.1.1 选题的依据

随着生活水平逐渐提高，电影成了丰富人们日常生活和精神世界的不可或缺的元素之一。电影市场在近些年发展得十分迅猛。据统计，2009年电影总票房为62亿元，此后每年保持40%左右的票房增长速度。2014年的年度总票房为296.3亿，而2015年的年度总票房增长到440亿，与前一年相比，涨幅高到48.73%，其中国产电影在总票房中的占比达到60%。面对如此巨大的票房市场，国内互联网三大巨头百度、腾讯和阿里巴巴在去年相继成立了自己的影视子集团，由此可见电影行业具有极大发展潜力。

传统的电影票房预测往往由各大院线工作人员（一般是各个院线的经理）根据历史数据和个人经验完成。但近年来，由于影片量级的快速增长、观影人数的急剧扩张以及宣传营销方式和社交网络的多样化发展，影响影片最终票房的因素变得越来越复杂，例如演员知名度、影片剧情、排片数、宣传力度、社交舆情、传播热度等[1]。根据个人经验和历史数据预估出的电影票房大多存在以下问题：第一，影响影片票房的因素的数据量庞大且数据差异性显著，个人不能准确地量化这些影响因素；第二，个人预测电影票房经常存在由于个人主观因素而导致非客观预测的现象；第三，近年来电影数量急剧增加，通过人工进行票房预测的工作量巨大，需要较多人力资源。上述问题最终会使得人工预测出的电影票房与票房实际值相差较大，即准确率比较低。而精准的电影票房预测值能在电影上映早期给电影的投资方和制片方带来营销和宣传上的数据支持，从而能在电影上映阶段获取更多观影票房。

随着大数据相关技术[2-3]和工具的快速发展，与之相关的统计学理论[4]和机器学习算法的应用越来越成熟，例如使用混合模型[5]进行推荐系统[6]的研发、利用朴素贝叶斯算法进行垃圾邮件的识别、利用随机森林算法对特征进行选择[7]、利用深度学习算法进行语音识别[8]等都是业界较为实际的应用。为了解决人工进行电影票房预测存在的种种问题，急需构建一个基于人工智能技术，例如数据挖掘和机器学习算法，并融合当下优秀的大数据处理工具，例如Scikit-learn以及Hadoop等进行电影票房预测的智能系统。

### 1.1.2 选题的意义

根据以上介绍的背景，伴随着网络通讯设备如智能手机、平板电脑、智能手表等的迅速普及，网络数据暴增，我们进入了大数据时代。大数据具有整体量级庞大、增长速度极快、类型繁杂等诸多特点。当前我国以及世界上其它部分国家已经发起了大数据战略。科研工作者和软件开发工程师可以从海量数据中提取出与实际需求相关的数据形式，并对这些数据进行数据清洗、数据转换、特征分析等操作得到目标数据，然后使用数据挖掘相关技术和算法对目标数据进行挖掘和分析，最后将这些数据应用于实际决策之中。在大数据相关技术不断创新的背景之下，业界出现了越来越多的高效大数据处理框架、工具和算法。如何充分利用这些大数据技术来解决电影票房预测问题是本文的意义所在。具体包含三个方面：第一，工程研究意义，GFS（Google File System）[9]和HDFS [10]等分布式文件系统在大数据技术的发展下应运而生，其中GFS由谷歌公司研发并在其内部广泛应用，而HDFS则在GFS原理的基础上研发并与Hadoop MapReduce[11]计算框架一同被业界广泛使用。Hadoop MapReduce作为一种分布式处理框架，拥有分布式计算能力强、容灾性强以及可扩展性好等优点，因此被大量应用于大规模离线计算任务。为了解决数据计算时重复I/O操作而产生的计算资源大量浪费的问题，基于内存计算的大数据框架的研究得到突破性进展，并诞生了以Spark[12]为代表的基于内存RDD（Resilient Distributed Datasets）算子的计算平台。Spark对计算过程中产生的结果进行缓存，而不是存储在磁盘上，因此具有计算效率高、I/O操作次数少等优点。此外，由于与影片相关的数据种类较多，且大部分数据无规律的存储在外部Web网站上，为了收集这些数据需要使用Python爬虫[13]技术进行大规模数据采集和存储。本文研究如何利用这些大数据工具进行电影票房数据的收集、存储、处理以及计算，具有工程研究意义。第二，理论研究意义，在大数据技术十分火热的同时，如何利用合适的理论算法挖掘出数据背后的价值是工业界和学术界研究的一个重要方向，与此相关的学科如数据挖掘[14]、机器学习[15]、模式识别等正受到计算机相关工作者越来越高的重视。在上一小节中，简要介绍了目前机器学习算法在业界的一些典型实际应用。当前，业界具有众多机器学习算法开源库，如Scikit-learn，它具有丰富的API接口、文档详细、使用简单等优点。本文研究如何利用机器学习算法进行电影相关数据的特征处理、数据计算以及准确票房预测，因此具有理论研究意义。第三，实际应用意义，本文的落脚点是设计一个智能化的电影票房预测系统，从而解决传统人工预测存在的的种种问题，并释放人力资源，具有很强的实际应用意义。

## 1.2 国内外文献资料综述

### 1.2.1电影票房预测研究进展

电影票房预测系统是一种智能的自动化系统，融合了数据收集、数据清洗、数据存储、数据计算、以及票房预测等较多功能，旨在弥补人工预测存在的诸多问题，提高票房预测的准确率和效率，为电影制片方和发行方提供准确的票房预测数据支持，从而进行更好的电影宣传和营销。

对电影票房进行预测起源于上世纪八十年代中后期，美国密歇根州立大学经济学家Barry Litman对当时最新的七百多部电影进行分析建模，并在此基础之上对影片票房进行预测[16]，他把决定电影最终票房的因素分成三类：影片创意、上映时间和影片营销方式，然后他根据这三类变量得出一个回归方程并进行票房预测。2006年，美国的两位学者Ramesh Sharda以及Dursun Delen使用神经网络算法对电影票房进行分档预测[17]，但是其准确率只有36.9%。此外，国外的Sitaram Asur等人在2010年使用推特的社交数据对票房进行了预测[18]。Berman等人提出使用反馈神经网络对票房预测盈利与否[19]，而不是预测票房实际值，Berman等人描述的神经网络只有一个隐藏层导致模型过于简单，实际使用价值不高。利用计算机技术进行电影票房预测真正进入人们的视野，是谷歌公司在2013年发表的一篇名为《Quantifying Movie Magic with Google Search》[20]的文章。从该文可知，谷歌主要使用影片排片数和相关搜索数据对票房进行预测。谷歌对外宣称他们预测的影片票房值在测试集上的判定系数R2为94%（其中R2取值区间为0到1，越大表示预测结果与实际票房越吻合，即预测效果越好），但他们一直没有公布对即将上映的电影的预测结果，因此无法对其准确率进行验证。

在国外对电影票房进行预测研究的同时，国内学术界和工业界也在这一领域展开了探索。百度投资子集团爱奇艺2014年对电影票房进行了预测，但其系统并未对外公开。而华影BRP近年来一直在对外公布单个影片的预测报告，但其系统的整体情况从未公开说明。重庆大学郑坚等2013年利用反馈神经网络对电影票房进行了分档预测[21]，其实验分类预测准确率高达93.69%，但其团队并未对票房的分布区间和具体值进行预测，因此不能满足实际业务的需求。国内的罗捷等人使用电影的评价数据，包括电影评分和电影评论数据对进口电影票房进行了预测[22]，但这种预测是基于后验数据而言，并不具备大规模实际生产要求，并且其实验对象只是针对国外电影，忽略了目前日益火爆的国内电影。

根据以上分析，电影票房预测经历了从人工预测到使用计算机技术进行预测这一过程。对电影票房预测系统进行进一步的研究十分必要，需要结合实际业务需求，并不断提高系统的预测准确率和预测效率。

### 1.2.2 大数据处理系统调研

近年来，与大数据技术相关的应用已经扩散到各个领域之中，例如医疗、金融、智能交通、智慧城市、信息安全等领域，将大数据与云计算相结合已成为了工业界的一个趋势。面对如此海量的数据，如何存储和高效地挖掘出人们关心的那一小部分数据是学术界和工业界一直在研究的问题。随着硬件计算性能地提升，诞生了许多不同的大数据处理系统。目前，大数据的处理任务主要分为批处理任务和实时处理任务，其中批处理任务是指实时性要求不高，可离线计算的任务。而实时处理任务则需要较高的响应速度。此外，大数据的存储主要基于键值对形式的海量分布式文件系统[23]。

大数据系统真正被人们所认识是谷歌公司2003年研发的GFS文件系统以及2004年研发的MapReduce分布式并行计算框架。但当时谷歌公司只公布了设计文档，并没有开源这两个系统。2006年，Hadoop发明者Doug Cutting根据谷歌公司的相关文档研发出了HDFS文件系统和MapReduce计算框架。Hadoop MapReduce主要用来处理离线大规模分布式并行计算任务，而HDFS则是一个能够对数据进行分布式并行存储的文件系统。在此基础之上，诸如Hbase，Cassandra[24]等诸多开源工具逐渐进入人们的视野。随着Hadoop的广泛应用，我国的互联网公司相继研发出自己的云计算应用平台。腾讯于2009年推出独立研发的云平台，目前其集群规模已达到上万台服务器；百度推出大数据处理引擎，支持TB级别数据查询，在此基础上研发的百度云盘是当下应用最广泛的云盘之一；另一互联网巨头阿里巴巴研发并推广的云平台ODPS[25]是国内先进的云平台之一，并已大规模投入商用和学术研究使用。阿里云在Sort Benchmark 2015年排序竞赛中对100TB数据进行排序仅仅花费了377秒，打破历史记录。除此之外，国内的七牛云，新浪云等也被越来越广泛的应用。

由于批处理系统在处理实时性要求高、响应速度要求快的任务上不具备优势，实时处理成了大数据技术实际生产中的一个重要研究问题。实时处理系统一般分为两类，一类是一次处理一批数据，这类任务一般吞吐量较高；另一类是一次处理一条数据，即流式计算，这类任务吞吐量较小。实时处理系统在谷歌公司于2010年推出Dremel[26]之后得到迅速发展。2011年推特公司对外发布其内部使用的Storm计算平台，其具有流式处理数据的功能。2009年诞生于伯克利大学的大数据计算系统Spark经过若干年的积累，目前已是工业界主流的实时处理平台之一，Spark的具有批处理、实时处理、交互式查询、图计算等功能，是一种基于内存RDD数据结构的计算框架。Spark对计算过程中产生的临时结果进行缓存，减少了中间计算过程的I/O开销，因此具有计算速度快的优点。由于Spark的计算是在内存中进行的，因此需要集群具有大容量的内存设备，而本文研究的电影票房预测系统大多数计算任务是离线计算形式，因此本系统没有采用Spark平台而使用了Hadoop计算平台。

### 1.2.3机器学习算法库调研

在大数据技术迅猛发展的同时，机器学习理论也越来越成熟。机机器学习是人工智能的核心，其主要内容是如何让计算机根据历史数据学习得到一个算法模型，并用该模型对实际数据进行计算得到一个预测结果。机器学习算法按照学习方式不同可以分为四大类：第一，监督式学习算法，该类算法的所使用的数据都有一个固定的标签，当算法学习完一个模型后对实际数据进行预测得到该数据的标签，主要应用包括回归和分类问题。第二，非监督式学习算法，该类算法的实际数据不具备明确的标签，因此算法需要学习数据潜在的关系，主要的应用包括关联规则以及聚类等问题。第三，半监督式学习，该类学习算法的训练数据往往部分有标签而剩余部分没有被标识。第四，强化学习，在这种学习方式下，算法模型需要立即对输入数据做出相应调整，使得奖励函数值最大，常见的算法包括Q-Learning[27]等。机器学习的算法根据实际应用又可以划分为回归算法、分类算法、聚类算法、关联算法等类别。

出于实际应用效率考虑，业界已经有众多机器学习算法库，这些算法库对各类算法进行封装并提供API接口，使用者只需要按照API接口要求进行调用并调整参数即可，调参是使用机器学习算法较为重要的一个步骤，好的参数组合往往能在一定程度上提升算法最终效果。Libsvm[28]是由国立台湾大学的Chin-Jen Lin博士于2001年使用C语言研发，Libsvm主要提供了非线性的的svm分类器算法，由于Libsvm具有高效、算法实现好和调参便捷等优点，其一经发布便受到广泛的关注和应用，Chin-Jen Lin博士于2008年研发并发布了用于大规模线性计算的Liblinear[29]。Liblinear提供了逻辑回归、线性SVM等线性算法接口，与上一代产品Libsvm一样具许多优点并提供了诸如C语言、Java语言以及Python等语言的API接口，遗憾的是实现的算法种类过少。Spark Mlib是Spark生态系统中的一个机器学习算法库，Spark Mlib实现了大多数常用的机器学习算法并与Spark RDD完美结合。随着深度学习理论的不断完善，2015年谷歌公司开源了其内部研发的机器学习算法库Tensorflow，与Theano、Caffe和Torch类似，它主要应用于人工神经网络和深度学习相关的问题，但需要大量的硬件计算资源和运行时间。由于Python具有开发速度快、高效以及开源库较多等优点，近年来，在数据处理上具有越来越显著的优势。基于Python研发的开源机器学习算法库Scikit-learn提供了常用机器学习算法和工具，具有算法实现好、文档齐全、算法种类多等优点，已经是该领域最受欢迎的机器学习算法库之一。由于Scikit-learn的这些优点，本文着重讨论基于Scikit-learn的电影票房预测系统的设计与实现。

## 1.3本人的主要工作

本人在基于Scikit-learn的电影票房预测系统的设计与开发中，参与了大量工作，具体包括如下内容：

1、分析系统需求与架构，调研实现系统应采用的理论和技术；

2、设计票房预测系统的特征工程，结合业务逻辑，从多维度挖掘特征；

3、实现数据抓取模块，从外部网站抓取与电影相关的数据；

4、实现自动化数据预处理和特征抽取模块；

5、实现票房预测模块，使用不同算法对历史数据训练得出的模型进行交叉

验证，得出每个算法的最好模型，并进行算法融合计算最终票房预测值；

6、对系统进行测试，确保系统的功能和性能符合要求；

7、分析系统存在的不足，并给出解决方案。

## 1.4 本文的组织结构

本论文共分为七个章节，每章主要讨论内容如下：

第一章为本文的绪论部分，主要介绍了本文的选题背景和研究意义，同时讨论了国内外相关领域的发展趋势，并对本人的主要工作进行了简要的说明；

第二章为相关理论原理与技术的介绍，主要介绍了整个系统开发过程中用到的工具，如Scikit-learn、Hadoop MapReduce计算框架、HDFS文件系统等，并分析了为什么要使用它们，同时详细描述了系统所需要使用的算法理论；

第三章对系统了行了需求分析，主要讨论了系统的功能性需求和非功能性需求，并对系统的可行性进行了讨论；

第四章阐述了系统的概要设计，主要分析了系统的整体架构和系统模块划分，并对每个模块的功能和数据接口设计进行了描述；

第五章主要讨论了票房预测系统的详细设计与实现，深入阐述了每个功能模块的具体技术细节以及相关算法；

第六章是本文的测试部分，主要介绍了系统的部署方法并对系统的功能和性能进行了测试分析；

第七章是工作的总结与展望，对已经完成的工作进行总结，同时提出了本系统进一步优化的方向。