**U**

**中国科学技术大学软件学院**

**《软件工程实验》项目环节**

**开题报告**

**项 目 名 称：基于机器学习算法的**

**影视数据分析系统**

**成 员 名 单： 邓高登 崔力辉 王世卿**

**指 导 教 师： 袁宇丹**

**工 程 领 域： 机器学习、自然语言处理**

**开 题 时 间： 2017年9月24日**

**中国科学技术大学软件学院**

**填表日期： 2017 年 10 月 25 日**

**一、 简况**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **名称** | **中文** | **基于机器学习算法的影视数据分析系统** | | | |
| **英文** | **Film and Television Data Analysis System Based on Machine Learning Algorithm** | | | |
| **项目组成员名单** | **姓名** | | **学号** | 项目中的分工 | 签 章 |
| **邓高登** | | **sa17225054** | **数据分析，词典生成** |  |
| **崔立辉** | | **sa17225045** | **数据分析，机器学习** |  |
| **王世卿** | | **sa17225363** | **数据分析，深度学习** |  |
|  | |  |  |  |
|  | |  |  |  |
| **中 英 文 摘 要** | **随着我国网民规模的增大，影视市场的快速壮大，视频网站的快速发展以及新型社交媒体的助力，我国的影视市场产生并积累了巨大的有效和有用的数据，在信息时代，这无疑是一笔巨大的财富。随着机器学习、自然语言处理等技术的发展，对于数据的价值挖掘和利用越来越受到人们的重视。**  **从近些年持续火爆的影视票房来看，我国影视市场无疑具有很大的发展前景。如何在影视市场中掌握先机，立于领先地位，是很多人特别是电影投资商关注的问题。针对这样的背景，本文提出了基于机器学习算法的影视数据分析系统，以期在对影视市场积累的数据进行充分利用的基础上，能够给予投资者投资影视市场的一些合理建议。**  **本工程实践从设计背景、需求分析、系统概要设计、系统详细设计、系统测试等角度对影视数据采集和分析系统进行了详细的叙述，并在实际工作中，完成了对该系统的设计和实现，从数据采集、数据预处理、数据分析和数据可视化四个方面实现了对影视数据价值的深度挖掘，可以给出能够屹立影视市场的良好建议。**  With the increase of the scale of Chinese netizens, the rapid growth of the film and television market, the rapid development of video websites and the help of the social media, the film and television market in China has produced and accumulated enormous effective and useful data, which, in the information age, is undoubtedly a huge wealth. With the development of the technology of machine learning and natural language processing, people pay more and more attention to mining and utilizing the value of data. Judging from the continuous hot film box office in the recent years, the film and television market in China undoubtedly has great development prospects. How to grasp the opportunity in the film and television market and take up the leading position, is the concern of a lot of people, especially film investors. Based on this background, this paper proposes a film and television data collecting and analyzing system, in order to be able to give investors valuable suggestions to invest in the film and television market, on the basis of making full use of the data accumulated by the film and television market.  In this project, we made a detailed description about the film and television data collecting and analyzing system from the views of the design background, the demand analysis, the system outline design, the system detailed design, system test, etc. And in the practical work, I completed the design and implementation of the system, and I implemented the deep mining of film and television data from four aspects including data collecting, data preprocessing, data analyzing and data visualizing, which could give good advice for people to make a success in the film and television market. | | | | |
| **主题词** | **主题词数量不多于三个，主题词之间空一格（英文用“/ ”分隔）** | | | | |
| **中文** | **机器学习 自然语言处理** | | | |
| **英文** | **Machine Learning/Natural Language Processing** | | | |

**二、选题依据**

1. **阐述该选题的研究意义，或工程设计的价值和意义，国内外概况和发展趋势，选题的先进性和实用性，技术难度及工作量。**

**随着传统的影业势力、新型社交媒体、在线售票与影评网站的快速发展，使得中国影视行业积累起了大量的有效和有用数据。例如，视频网站比如爱奇艺，优酷，腾讯视频；影评网站比如豆瓣，时光网；新近的电影票购买和电影评论网站如猫眼，格瓦拉等都积累的大量的用户，也给影视产业的发展提供了大量可以利用的数据：在视频网站如优酷网，可以获得影视剧的播放量，地域分布等比较直观的数据；在猫眼，时光网上可以获得电影的每日票房；在豆瓣等网站上可以获取大量的影评等等。**

**利用这些数据，我们可以探索隐藏在其背后的宝藏。最常见的例子，从普通用户的角度来讲，从电影票房是否火爆可以看出一部的电影是否值得买票去看；而从影视投资的角度，我们可以获得对于从影视制作环节的IP评估、剧本、编剧、导演、演员等主创人员的确认，到营销环节的目标观众偏爱分析、适播平台分析，再到近年来兴起的品牌植入效果等方面的指导信息。**

**举例来说，对于题材评估，如果计划投拍一部影视剧，则需要了解该项目的市场收益及市场风险。项目的题材是否符合当下的市场需求？项目的类型是否在当下较火？以电影市场为例，我们在面对这些问题的时候，可以通过分析各种类型、题材的票房，得到该电影类型是否比较受欢迎，题材是否符合市场的需求；也可以对比历史收益最高的电影，效费比最高的电影，好评度最高的电影分别进行统计和分析，从而判断热门题材和冷门题材。**

**对于班底，通过对海量数据的挖掘统计分析，选择合理的算法，对每个导演、演员、编剧等制作人员及其班底的资质、各种题材的经验、所拍摄过的电视剧的收视率进行全面的评价，通过数据展示班底成员所参与或拍摄过的电视剧、每个电视剧的基本资料及收视信息进行分析，给投资方提供一份相对客观、用数据说话的评估结果，包括酬劳是否合理，从而选择一套合适的班底。**

**本工程实践的技术部分主要分为四大部分，一是数据采集，主要通过爬虫技术进行了实现；二是数据处理，处理的对象是采集的数据中的无效和无用数据；三是数据分析，通过自然语言处理中的聚类、情感分析等方法对于挖掘出数据中所蕴含的有价值的信息；四是数据可视化，通过前端技术将分析得到的数据以图表的方式进行展示。**

1. **参考文献（参见：中国科学技术大学研究生学位论文撰写规范）**

**[1] Vural, A. Gural, B. Barla Cambazoglu, and Pinar Karagoz. "Sentiment-focused web crawling." ACM Transactions on the Web (TWEB) 8.4 (2014): 22.**

**[2] 王景中, and 邱铜相. "基于 TF-IDF 改进算法的聚焦主题网络爬虫." 计算机应用 35.10 (2015): 2901-2904.**

**[3] 周胜臣, et al. "中文微博情感分析研究综述." 计算机应用与软件 30.3 (2013): 161-164**

**三、课题内容及具体方案**

|  |
| --- |
| **拟采用的设计方法，技术路线，实验方案的可行性分析。（需注明由哪位成员完成）**  **本工程实践的技术部分主要分为四大部分：一是数据采集，主要通过爬虫技术进行了实现；二是数据处理，处理的对象是采集的数据中的无效和无用数据；三是数据分析，通过自然语言处理中的聚类、情感分析等方法对于挖掘出数据中所蕴含的有价值的信息；四是数据可视化，通过前端技术将分析得到的数据以图表的方式进行展示。**  **第一部分数据采集，需要使用爬虫进行数据的采集。其原理如下文所述。传统爬虫从一个或若干初始网页的URL开始，获得初始网页上的URL，在抓取网页的过程中，不断从当前页面上抽取新的URL放入队列,直到满足系统的一定停止条件。聚焦爬虫的工作流程较为复杂，需要根据一定的网页分析算法过滤与主题无关的链接，保留有用的链接并将其放入等待抓取的URL队列。然后，它将根据一定的搜索策略从队列中选择下一步要抓取的网页URL，并重复上述过程，直到达到系统的某一条件时停止。另外，所有被爬虫抓取的网页将会被系统存贮，进行一定的分析、过滤，并建立索引，以便之后的查询和检索；所以一个完整的爬虫一般会包含如下三个模块：(1) 网络请求模块 (2)爬取流程控制模块 (3)内容分析提取模块。（此部分由三人完成）**  **第二部分数据处理，使用网络爬虫采集的信息，需要分类整理存储到数据库中，在从网络上获取数据的时候，为了保证取得想要的数据而舍弃不需要的数据，需要对来自网络的数据进行过滤，这样才能从数据采集的源头保证数据的质量；而对于已经采集的需要存储在数据库中的数据，清洗也是不可避免的，比如可能出现需要保存的数据中有奇怪的符号而无法存储到数据库中；同时，存储在数据库中的数据还应该按照后面数据分析部分的要求，进一步清晰数据，提高数据质量，进而提升数据分析的结果。（此部分由崔力辉完成）**  **第三部分数据分析，主要使用机器学习算法，对抓取到的影视评论通过自然语言处理技术进行情感分析。情感分析，又称倾向性分析，意见抽取，意见挖掘，情感挖掘，主观分析，它是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。现阶段主要的情感分析方法主要有三类：（1）基于词典的方法（2）基于机器学习的方法：这一块涉及到常用的分类算法模型，如朴素贝叶斯、逻辑斯谛回归、支持向量机等（3）基于深度学习的方法，这一块主要利用CNN、RNN(LSTM)等神经网络结构，通过SGD（Stochastic Gradient Descent）算法在训练数据上利用GPU进行训练。因为模型本身Embedding有包含语义信息和网络结构保留了词序信息，大量数据的训练结果会使得效果上是要好传统的机器学习。（此部分由三人共同完成，邓高登负责词典，崔力辉负责机器学习，王世卿负责深度学习）**  **第四部分数据可视化，数据可视化可以提高决策的效率，将抽象的非结构化的数据进行可视化处理，使得在有限的展现空间中以直观的方式传达大量的抽象信息，对于决策者，可以避免决策的低效，减少金钱和时间上的损失，使决策更科学。为了实现影视数据采集和分析系统的可视化效果，需要使用恰当的前端技术，根据需求，将用到的一些具有代表性的前端技术列举如下：Node.js、Handlebars、ECharts等。（此部分由三人共同完成）** |

**四、工作进度的大致安排**

|  |  |
| --- | --- |
| 应包括文献调研，理论分析，数值计算，理论分析，软硬件设计，仪器设备的研制和调试，撰写论文等，要给出各个阶段的成果形式。  **2017年11月——2017年12月**  **数据采集，网络爬虫编写、改进、调优、封装，**  **文献摘抄记录**  **2017年12月——2018年1月**  **数据处理，数据清洗、数据存储，**  **采用命名实体识别的方法，结合语义角色标注等方法，过滤无效评论**  **学习笔记**  **2018年1月——2018年3月**  **数据分析**  **机器学习常见方法的学习、几种分类算法的学习、深度学习算法学习**  **学习笔记、算法实现代码**  **2018年3月——2018年4月**  **数据可视化**  **主要代码的编写、测试、评估、调参**  **实现代码、代码文档**  **2018年4月——2018年5月**  **结题报告的编写、系统的最终调优**  **完整的推荐系统、结题报告** | |
| **预期成果及特色**  **项目旨在完成一个基于Spark大数据平台的机器学习电影推荐系统，通过不同的算法为用户推荐电影，并试图找出最优解。** | |
| **导师意见（对选题和工作过程及成果进行说明，并给出成绩。）**  **导师签名：**  **年 月 日** | |
| **答辩小组意见**  **答辩组长签名：**  **年 月 日** | |