北京化工大学学位论文原创性声明

本人郑重声明： 所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

作者签名：徐振波 日期：2016年5月30日

关于论文使用授权的说明

学位论文作者完全了解北京化工大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属北京化工大学。学校有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。

□论文暂不公开（或保密）注释：本学位论文属于暂不公开（或保密）范围，在 X 年解密后适用本授权书。

□非暂不公开（或保密）论文注释：本学位论文不属于暂不公开（或保密）范围，适用本授权书。

作者签名：徐振波 日期：2016年5月30日

导师签名：李健 日期：2016年5月30日

万方数据

学位论文数据集

Y3094367

中图分类号	C931.1	学科分类号	630.5035
论文编号	1001020160869	密　　级	公开
学位授予单位代码	10010	学位授予单位名称	北京化工大学
作者姓名	徐振敬	学　　号	2013200869
获学位专业名称	管理科学与工程	获学位专业代码	087100
课题来源	国家自然科学基金研究	研究方向	石油预测
论文题目	基于情感分析的国际原油价格走势预测研究		
关 键 词	情感分析 文本挖掘 石油走势预测 人工智能 格兰杰因果检验		
论文答辩日期	2016 年 5 月 22 日	* 论 文 类 型	应用研究

学位论文评阅及答辩委员会情况

	姓名	职称	工作单位	学科专长
指导教师	李健	教授	北京化工大学	物流与供应链管理
评阅人 1	李想	教授	北京化工大学	管理科学与工程
评阅人 2	马骏	教授	对外经贸大学	管理科学与工程
评阅人 3				
评阅人 4				
评阅人 5				
答辩委员会主席	方勇	教授	北京化工大学	技术经济
答辩委员 1	方勇	教授	北京化工大学	技术经济
答辩委员 2	刘斌	副教授	北京化工大学	技术经济
答辩委员 3	任继勤	副教授	北京化工大学	技术经济
答辩委员 4	吴卫红	副教授	北京化工大学	技术经济
答辩委员 5	张爱美	副教授	北京化工大学	技术经济

注：一. 论文类型：1. 基础研究　2. 应用研究　3. 开发研究　4. 其它

二. 中图分类号在《中国图书资料分类法》查询。

三. 学科分类号在中华人民共和国国家标准（GB/T 13745-9）《学科分类与代码》中查询。

四. 论文编号由单位代码和年份及学号的后四位组成。

万方数据

基于情感分析的国际原油价格走势预测研究

摘要

石油价格的增长或者下跌对于世界经济的发展会产生重大的影响，因此，石油价格的预测显得尤为重要。但是，由于石油这种商品的复杂性和不规律性，使得石油价格预测成为一个非常困难的问题。石油价格基本是由石油出口国的供给和石油进口国的需求平衡共同决定的，但是由政治事件或者经济因素导致的供给的不规律性往往会导致石油价格的不规律性。在这种情况下，石油价格上涨或者下跌的预测对于决策者就显得非常有价值。同时，随着互联网和大数据技术的发展，随之产生了大量新闻数据。一些新闻内容代表了对金融市场未来趋势的实时评估，新闻的内容将会影响金融投资者的投资行为，进而影响金融市场。如果能充分利用这些数据，那么将有助于石油价格趋势的预测。

在这种背景下，本文提出了基于情感分析的国际原油价格走势预测模型。该模型主要基于石油相关新闻的分析，通过采用领域关键词词典的方法，得到新闻的情感序列，再通过格兰杰因果检验的方法，得到情感序列和石油价格序列的相关性和滞后期，最后通过机器学习的方法(支持向量机、决策树、逻辑回归和神经网络)预测石油价格的走势。同时，为了验证新闻情感对于石油价测具有预测能力，选取了美国西得克萨斯轻质原油

万方数据

（WTI）和路透社原油新闻作为研究对象进行案例分析。结果表明，新闻情感和原油价格之间确实存在着格兰杰因果关系，即新闻情感的变化会引起石油价格的变化，这说明新闻情感对于石油价格的走势具有预测能力。实验结果还表明，对于大部分预测模型而言，新闻情感的引入一般能极大地提高石油价格走势的预测准确率。

关键词：情感分析，文本挖掘，石油走势预测，人工智能，格兰杰因果检验

万方数据

FORECASTING OIL PRICE TRENDS WITH NEWS SENTIMENT

Abstract

The growth or decline in oil prices will have a significant economic impact on the whole world. Under such background, efficient and accurate predictions for crude oil price are critical for a stable economic development. Despite such attempts, oil price prediction has remained a difficult problem due to its complexity and irregularity. The price of oil is basically determined by balancing the amount of oil the net oil-exporting countries can supply with the demands of the net importing countries, but the irregularity is caused more by the shocks on the supply-side, which may be political disputes or sudden changes in external economic factors. In such cases, a precise prediction of the values of the oil price will be difficult to obtain. However, a rough prediction of the upward and downward changes of the price can still be helpful for decision making. With the rapid development of the Internet and big data

万方数据

technologies, a variety of news data to be processed continues to witness a quick increase. In fact, content of news information represents the real-time evaluation on future trend of financial markets. The mainstream news and information will be reflected in the emotional final of the investment behavior of financial practitioners, and thus affect the stock market. If we can take advantage of this big data on the Internet, we can analyze and forecast the movement direction in crude oil prices. Generally speaking, this paper introduce news sentiment, which is extracted based on a dictionary-based approach, into crude oil price trend forecasting and employ several powerful machine algorithms (i.e., SVM, DT, LogR and BP) to verify the predictive power of news sentiment. Also, we use a Granger causality analysis to investigate whether news sentiment correlate with changes in crude oil price and to determine the predictive lag order. For illustration and verification purposes, the crude oil future price in West Texas Intermediate (WTI) market and the news of crude oil market are collected from the international news agency Thomson Reuters are taken as sample data. Empirical results of Granger causality analysis statistically show that, news sentiment has significant Granger causality relation with crude oil price and the variations of news sentiment more likely lead to changes in crude oil price, which indicate the predictive power of news sentiment. Empirical results of trend prediction

万方数据

show that adding news sentiment have significant effect on prediction accuracy compared to using only historical crude oil price values.

KEY WORDS: sentiment analysis, text mining, prediction of movement of oil price, artificial intelligence, granger causality investigation

万方数据

万方数据

目录

万方数据

万方数据

Content

万方数据

万方数据

第一章 绪论

1.1 课题的研究背景

从1983年以来，石油已经被当做世界上最重要的商品之一[1]。由于石油供给和需求的巨大变化，石油价格变化也非常之大。尽管如此，作为世界上大多数国家的能源需求，石油的发展速度非常快。石油，无论对于石油出口国还是输出国，都极大地促进了当地的经济的发展[2]。石油及其衍生品对于我们这个世界来说都具有重要意义，石油经过分解得到许多产品对于世界经济发展都具有重要作用。

近些年来，化工行业和交通运输业的发展，导致石油的需求急速上升，因此全世界的石油贸易迅速增加。2007年，全球总共成交13.85亿石油期货合约，其中每份占1000桶石油。除此之外，石油输出国组织（OPEC）预测石油需求到2035年将会增加2亿通。这些贸易的增加带动了石油运输产业的发展，当然也促进了世界各地经济的发展。但是最近石油价格急剧震荡，石油价格[3]从2003年到2008年逐渐增长，到2008年时达到145美元/每桶，到2015年大约又跌了100美元。石油价格的震荡，无论是对于石油输出国，还是石油进口国的经济都有巨大的影响。对于石油纯进口国来说，高油价将会导致国民收入的下降，工资下降，高失业率，税收减少，高利率等等；而对于石油纯进口国来说，则是相反的，石油价格上涨将直接提高石油出口国的收入。从长期来看，其中一部分收入将会由于石油进口国的经济衰退导致的需求疲软而减少。石油价格上升的越大，对于全球经济的影响越大。而低油价对于石油出口国来说，将会导致石油产量迅速下降，同时可能带来石油生产国的债务危机，高失业率，信用危机，经济萧条的等负面作用；对于石油进口国来说，低油价意味着制造业成本的降低，促进本国经济发展。

我国作为世界上最大的发展中国家，世界第二大经济体，现如今已经是世界上最大的石油消费国。2015年中国石油进口达3.34亿吨，同比增长8.8%，2015年石油消费5.43亿吨，对外依存度首次突破60%，石油价格的变化对于我国经济来说更会带来巨大变化。

在此背景下，无论是国内学者还是国外学者都积极投入到了石油价格预测研究中。伴随着石油价格的不稳定变化和世界经济的放缓，从石油价格变化行为理论研究，到石油价格和经济指标相关性分析理论研究，再到石油价格预测理论研究，国内外学者涌现了一大批优秀的研究成果。其中，石油价格变化行为理论，旨在研究石油价格的影响因素以及各因素的相互作用，这对于弄清石油价格变化的机理具有重要意义。而

万方数据

石油价格和经济指标相关性分析主要研究石油价格同经济发展指标的相关关系，这些理论对于研究石油价格对于世界经济的影响具有重要作用。对于石油价格预测领域来说，主要指一系列的石油价格预测方法的研究。

随着互联网和信息科技的发展，随之产生了大量数据[4-6]。有效的管理和处理这些大规模的数据已然成为了一个巨大的挑战。近些年来，大数据已经吸引了学术界，工业界和政府部分的关注。特别的对于金融市场，一大批研究学者已经基于大数据的处理来研究金融市场的运作规律。根据有效市场的假设，金融信息对与金融市场的波动性具有重要影响。事实上，新闻内容代表了对金融市场未来趋势的实时评估。新闻的内容将会影响金融投资者的投资行为，进而影响金融市场。石油，作为一种特殊的商品，其影响因素，例如经济，军事和政治因素，自然灾害，投机和供给需求，都汇集在互联网上。所有的这些著名网站，产生大量数据。如果能好好利用这些数据，那么对于石油价格趋势的预测将产生重要的影响。

1.2 课题的研究目的与意义

石油价格的走势决定着国家经济的发展速度和发展规模，是国家发展，企业发展和社会进步的重要影响因素。准确油价走势的预测，对于石油进口国来说，可以依据油价走势科学合理的制定石油生产计划，是维护国内经济稳定和保障人民生活稳定的重要保证，而对于石油消费国来说，石油资源是国家和社会的重要的能源，准确的油价走势预测，可以使国家和企业合理的制定石油进口计划，合理规避风险，企业避免在油价急剧震荡中蒙受损失，也是国家安全稳定的重要保障，。

石油价格预测仍然是一个困难的问题，这主要是由于石油这种商品的复杂性和不规律性。复杂性主要由于许多全球经济或者国家经济因素的交错影响，这些影响是极其显著而却难以量化的。另一方面，石油价格的暴涨暴跌在过去曾经发生过。例如在1973 年，1978 年和 2008 年，出现过石油价格的暴跌，而在 1986 年和 1998 年，出现过价格的暴涨。石油价格基本是由石油出口国的供给和石油进口国的需求平衡共同决定的，但是往往由政治事件或者经济因素导致的供给的不规律性将会导致石油价格的不规律性。在这种情况下，石油价格的预测将会非常困难，然而石油价格上涨或者下跌的预测对于决策者就显得非常有价值。

随着互联网和大数据技术的发展，基于大数据技术的市场预测已经变得越来越受关注[4-6]。如何有效的管理和使用这些大规模的数据辅助预测已然成为了一个重要的研究课题。近些年来，学术界，工业界和政府已经开始关注并使用大数据解决传统问题。

万方数据

特别的对于金融市场，一大批研究学者已经基于大数据的处理来研究金融市场的运作规律。根据有效市场的假设，金融信息对与金融市场的波动性具有重要影响。事实上，新闻内容代表了对金融市场未来趋势的实时评估。新闻的内容将会影响金融投资者的投资行为，进而影响金融市场。而对于石油市场来说，石油的影响因素众多，像经济、军事和政治因素、自然灾害、投机和供给需求等因素，以不同的载体形式（如新闻、评论、微博或者博客）都汇集在互联网上，他们包含了影响石油市场运作的重要信息，而这是被传统的石油价格预测模型忽略的，因此本文旨在通过提取互联网中的重要的信息（知识）来辅助石油价格的预测。

新闻的文本内容提供除了关于事件事实的报道，还包括语言的语调或者情绪。而这些语调或者情绪反映了市场对石油价格走势的理解和判断，而这些新闻通常来源于市场重要的专家或者学者，因此这些语调代表了市场参与人员对石油市场的重要的态度，因此借助于这些情绪可以对于石油价格走势做出更准确的判断。因此本文考虑把新闻情绪引入到石油价格走势的预测中。

1.3 本文的研究内容

本论文的研究主线是在基于文本挖掘的国际原油价格预测的框架下，在大数据的背景下，利用石油市场新闻的情感倾向性，利用格兰杰因果检验判定新闻情感和石油价格的相关关系，最后利用人工智能模型预测石油价格的走势变化。具体来讲如下，鉴于世界上三分之二的能源需求来自于石油和天然气[67]，石油价格的上涨或者下降对与石油进口国和石油出口国具有重要影响。石油的价格已经成了全球或者国家经济表现的重要决定指标。石油价格的增长或者下跌对于全世界都会产生重大的经济影响[68]。因此石油价格的状态成了经济学家和政治家们关注的重点，但同时由于石油价格的影响因素复杂繁多，使得石油价格预测成为一个非常困难的问题，石油价格基本是由石油出口国的供给和石油进口国的需求平衡共同决定的，但是往往由政治事件或者经济因素导致的供给的不规律性将会导致石油价格的不规律性。在这种情况下，石油价格的预测将会非常困难，然而石油价格上涨或者下跌的预测对于决策者就显得非常有价值。并且，随着互联网和大数据技术的发展，基于大数据的市场预测越来越受到关注[4-6]。 有效的管理和处理这些大规模的数据已然成为了一个巨大的挑战。而对于石油市场来说，石油的影响因素众多，像经济、军事和政治因素、自然灾害、投机和供给需求等因素，以不同的载体形式（如新闻、评论、微博或者博客）都汇集在互联网上，他们包含了影响石油市场运作的重要信息，而这是被传统的石油价格预测模型忽略的，

万方数据

因此本文旨在通过提取互联网中的重要的信息（知识）来辅助石油价格的预测。

在这种背景下，本文提出了基于情感分析的国际原油价格走势预测模型，该模型主要基于路透石油相关新闻的分析，通过采用领域关键词词典的方法，得到新闻的情感序列，再通过格兰杰因果检验的方法，得到情感序列和石油价格序列的相关性和滞后期，最后通过机器学习的方法预测石油价格的走势。本文技术路线图如图1-1所示，主要研究内容分为五个部分：第一章为绪论，主要介绍本文的研究背景，分析研究的目的与意义以及本文的研究内容；第二章为文献综述，主要对石油价格预测、文本挖掘和情感分析在研究内容和研究方法上进行分析总结，指出了现有的研究不足和本文的创新之处；第三章首先描述了情感分析和格兰杰因果检验的理论，其次依次介绍了石油价格及新闻数据收集、爬虫程序的设计以及新闻数据的分析，最后进行了格兰杰因果分析；第四章首先介绍了各种预测方法的理论，其次对基于情感分析的石油价格走势预测模型进行了探讨，并基于获取的新闻数据进行了实证研究；最后一部分对本文的研究成果进行了总结与展望。

万方数据

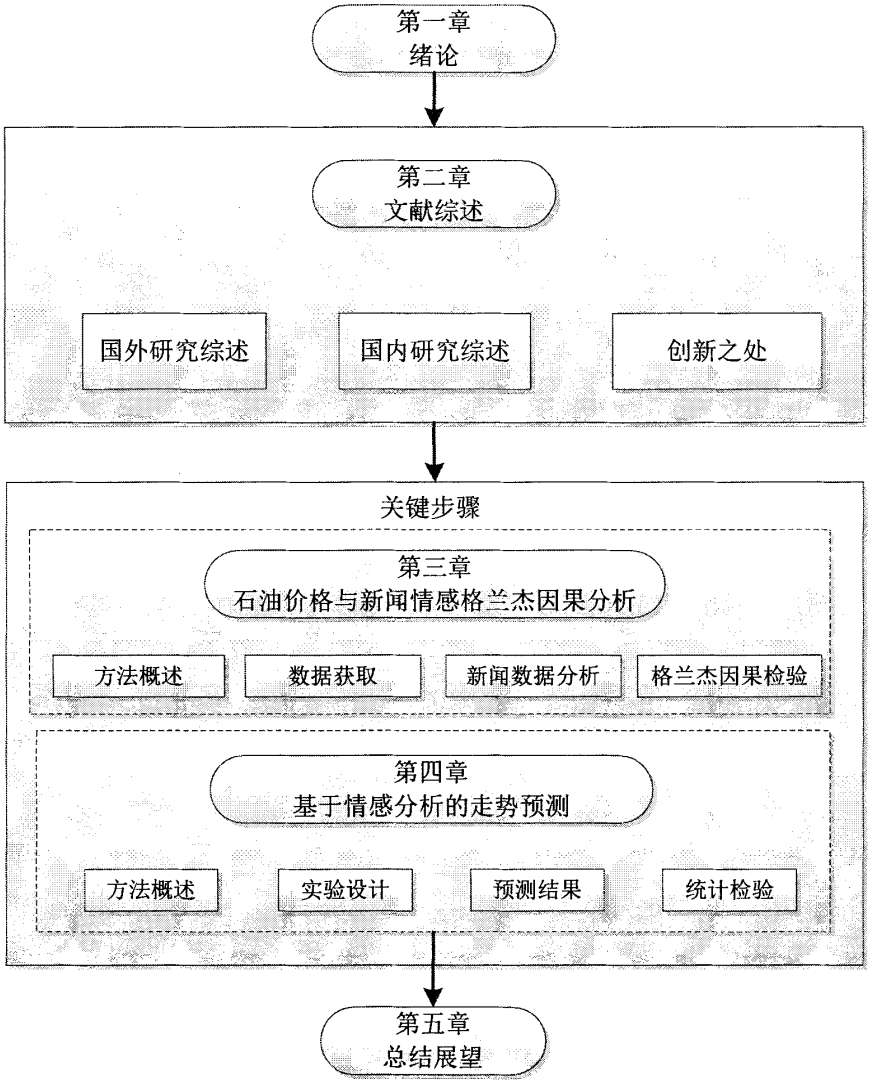

图 1-1. 论文框架图

Figure 1-1 Framework of the paper

万方数据

万方数据

第二章 文献综述

针对于本课题研究内容，主要从以下几个部分分别从国内外角度进行研究现状的分析与讨论。

其一，石油价格预测的研究内容和主要方法。其二，文本挖掘在市场预测领域的主要研究内容和研究方法。其三，情感分析的应用领域和主要研究方法。

2.1 国外研究综述

国外学者在石油价格预测领域研究起步较早，已经初步形成了一整套完整的预测方法理论。

其中，最常见的就是基于统计的时间序列方法[7-8]，这些方法认为石油价格特征由趋势、季节因素、周期因素和随机误差项所刻画。ARIMA 模型[9,76,77]，ECM 模型[10-12]和 VAR 模型[13]是其中的重要预测方法。ARIMA 模型，作为一种典型的传统的计量方法，已经被很多学者作为石油价格预测的对比模型[14-15]。Lanza 等人[10]用 ECM 方法研究了原油价格和石油产品关系。

另外一些研究方法假定石油价格和一些经济指标具有重大关系。Oladosu[16]采用 EMD 经验模态分解的方法研究美国 GDP 和原油价格的关系，认为两者具有负相关关系。King 等人[17]发现政治事件和经济新闻，同石油的供给和需求一样，对于石油价格的影响具有重要作用。

还有最近新兴起的一些方法主张采用机器学习的方法来预测石油价格。例如 Kulkarni 和 Haidar[18]基于一个多层感知的神经网络预测石油价格。Khashman 和 Nwulu[19]采用 SVR 模型预测石油价格，同时取得了较高的准确率。Yu 等人[20]提出来一个基于压缩感知和人工智能算法结合的石油价格预测模型，该模型使用压缩感知来减少原始石油序列的噪声，然后再使用一系列人工智能模型来预测石油价格，结果显示该模型具有很好的预测精度和鲁棒性。可以看出，最近几年，已经有越来越多的学者使用人工智能模型研究石油价格预测模型，这些都源于人工智能模型对非线性数据超强的拟合能力，因此本文在石油价格走势中也引入来人工智能模型。

国外学者在文本挖掘技术应用于市场预测领域已经取得了许多优秀的成果，而且其中不少已经商业化。而在文本挖掘技术在市场预测领域的研究内容本文主要从三个方面概述，输入数据，预处理和机器学习方法。输入数据就是作为整个预测框架的输

万方数据

入，而预测值作为真个系统的输出。

首先，输入数据主要有包括两种网上文本数据和市场数据。文本数据大部分来源于新闻网站，例如华尔街日报[21]、金融时报[22]、路透社[23]、道琼斯指数[24]、福布斯[25]和雅虎[26]。文本数据的类型主要是一般新闻和金融新闻，而金融新闻又占了一大部分，这主要是因为金融新闻相比于一般新闻数据更纯净，噪音更少。拿新闻来说，更多的研究学者喜欢使用新闻标题作为研究语料，因为新闻标题更直接，更切中要点，且明确表达了主题[27-28]。还有一些学者使用网上评论版块（留言板）的数据。但是现在越来越多的学者开始使用社交媒体像推特和微博的信息作为研究语料，并且他们其中一些人认为这些社交媒体的数据作为预测信息，才能更准确的反映投资者的态度。另外，还有一些研究学者把公司报告和行业报告作为研究文本。使用这类文本最大的好处就是，它们有着非常强的预期安排规律，从这个角度讲，这些报告对未来有一定的预测能力，因为它们的制作是考虑到未来市场可能的变化而制定的[24,26]。

其次，在文本挖掘技术应用于市场预测中，市场数据主要有市场价格和各种经济指数。这些数据的主要用来训练预测模型。过去几年的研究主要集中于股市预测，像股市指数预测如道琼斯工业指数[21,22,29]、美国纳斯达克指数[25]、斯坦利摩根高科技指数[3]、印度 Sensex 指数[30]、S&P500 指数[31]、或者股票价格预测如 BHP Billiton[32]、苹果、谷歌、微软和亚马逊[33]和一系列其他公司的股价。

在文本挖掘的预处理当中，主要有特征选择，特征降维和特征表示这方面的研究内容。

首先是特征选择，特征选择对于文本挖掘是非常关键的，因为只有重要的、关键的特征对于学习训练模型才是有意义的。常用的特征选择方法主要有词袋法[22,23,28]、名词词组和命名实体[31,34]、主题模型（Latent Dirichlet Allocationí）[30,35]和 n-gram 模型[3,36]。其中词袋法是最常见的方法，该方法把文本分解成词，每一个词表示为一个特征，注意这种方法是不考虑词出现的顺序的，并且词组的共现也是忽略的。Schumaker 等人[26]和 Schumaker and Chen[31]提出了使用名词词组和命名实体来选择特征，前者基于预先定义好的名词词典和语法规则提出名词词组，而命名实体识别系统也是基于名词和名词短语，识别出文本中的时间、地点、金钱、组织和人名等。另外一个较少使用但是很有趣的一个模型是主题模型，该模型把词转换为概念，把这些概念作为文本特征。还有一些文章使用 n-gram 模型，一个 n-gram 指的是文本中连续 n 个词组成的序列，有的还会考虑句法信息。

限制特征的维度是极其重要的，因为特征维度的增加将会使得分类任务或者聚类任务变得非常复杂，因此解决此问题的一个重要的方法就是特征降维。一些文献中常

万方数据

用的降维方法如下， Zhai 等人[32]通过选择前 30 个概念作为特征来达到降维的目的；2004 年 Mittermayer[37]也是选择前 1000 个词组作为特征，其实通常最常使用的是通过设定最小出现频率值来选择特征[31, 36]。另外一种常用的降维方法是使用预先定义的词典，替换某些词组。这些词典通常由市场专家制定[22,28]。

在最小特征维度确定后，每一个特征需要表示成一个数值，以至于机器学习算法可以去学习。这些数值可以理解为得分或者权重。常用的五个比较流行的表示方法，分别是布尔表示，信息增益（Information Gain），卡方统计值（Chi-square Statistics），DF（Document Frequency），TF-IDF(Term Frequency-Inverse Document Frequency)。布尔表示是最常用的特征表示方法，如果词袋中的一个词出现了，那个这个特征值就表示 1，否则为 0[22,26,30]。另外一个常用的表示方法是 TF-IDF[23,28,34,53]。TF-IDF 计算分值时，主要综合考虑了 TF 和 IDF 的思想，其主要思想是对那些在某类文章中出现频率高，而在其他文章中出现频率低的词汇赋予更高的分值，因为这些词更具有分类能力。

在经过预处理和特征表示之后，机器学习算法开始学习分类模式。这里本文主要列举一下常用的分类模型，并不对它们进行对比。基本上所有的分类算法都是通过对输入数据处理分析得到市场移动到走势，是上涨，下跌还是保持不动。常见的分类模型有支持向量机（Support Vector Machine），回归算法，贝叶斯（Naïve Bayes），决策树（Decision Rules or Trees）以及它们的组合算法。

新闻的文本内容提供除了关于事件事实的报道，还包括语言的语调或者情绪。通过情感分析的技术手段可以从这些文本内容中提取到关于报道的情绪。情感分析，也可以叫做态度挖掘，主要是挖掘出文本内容是正向情感还是负向情感。

在金融市场，信息情感是一种反映投资者和交易者的情感和观点的重要指标。前人的研究也已经证实了文本情感和股票市场波动的紧密相关性。Devitt 和 Ahmad (2007)[39]使用金融评论的情感倾向性预测未来金融趋势。Das 和 Chen (2007)[3]使用线性回归的方式证明股票指数和在线情感分析有高度的正相关关系，但是对于一个单独的公司，这种关系并不明显。Johan, Bollen 等人(2011)[29]使用格兰杰因果检验的方式验证情感状态和 DJIA 指数的关系，并使用情感去预测 DJIA 指数，结果显示可以明显提高预测效果。但是，在石油市场领域，基于新闻的情感分析和石油价格的波动研究还非常的少，但是鲜有的几个研究已经证实了情感分析对于石油价格的重要作用。Lechthaler 和 Leinert, (2012)[40]使用 VAR 模型研究 2003 年以后的全球原油市场的价格的动态波动，表明新闻情感对于石油价格的变化具有重要的影响。Feuerriegel 等人 (2014)[41]使用新闻的情感预测投机泡沫，表明新闻情感显著的影响的石油价格的收益

万方数据

变化。Alfano 等人(2014)[42]通过对新闻情感和石油时间序列的分解成分的回归分析，认为新闻情感不仅对于石油的噪声残差有影响，而且对于基本价格趋势也有影响。

信息系统方面的研究已经发展出了几种情感分析的方法。例如，Pang 和 Lee (2008)[43]提供了一个基于领域的情感分析研究的调查。调查显示，在金融市场领域，最近的研究[38,44,45]）主要集中于股票市场预测。金融文本挖掘方法主要依赖于基于词典的方式[46-50]。这种基于词典的方式通过计算在给定词典中出现的预先定义的正向词和负向词的频率。机器学习是另外一种分析方法[31,37,51,52]，但是这种方法的一个缺点就是可能导致过度拟合。

2.2 国内研究综述

近些年来，由于世界经济一体化趋势的加快以及全球经济形势的复杂变化趋势，我国学者在石油价格预测领域也取得了很多的研究成果。最近几年石油价格的预测方法也主要集中在基于统计的时间序列方法和人工智能方法。在基于统计的时间序列预测方法中，侯璐（2009）[54]应用 ARIMA 模型来进行石油价格的短期预测。肖龙阶和仲伟俊（2009）[55]也是通过 ARIMA 模型研究我国石油价格的预测问题。丁静之等人（2008）[56]主要通过定性分析与定量分析相结合的方法预测石油价格的走势。而在人工智能方法上，越来越多的学者开始使用像神经网络，支持向量机和贝叶斯网络来预测石油价格。例如，贾振华等人（2011）[59]提出了一种使用神经网络预测石油价格的方法；王欣冉等人（2011）[60]提出了基于小波包和贝叶斯最小二乘支持向量机相结合的石油价格预测方法。

在文本挖掘技术在市场预测方面的研究从 2008 年之后，已经有越来越的学者开始注重起来。如韩春和田大钢（2008）[61]对于股票市场开始使用文本挖掘的技术挖掘潜在有价值的信息，并同股市行情对比做相关性分析，但是并没有把这些挖掘到信息应用到股市的预测中去。赵丽丽等人（2012）[62]通过对财经新闻的的分析来量化对中国股市的影响。陈茜和连婉琳（2015）[63]通过文本挖掘技术对互联网中的股票新闻做情感的分类。最近，同样有国内的学者开始微博等社交媒体的数据应用到股票预测中去。例如，张世军等人[64]提出了基于网络舆情的，结合支持向量机的股票价格预测模型；黄润鹏等人（2015）[65]提出了基于微博情绪信息的股票价格预测模型。

基于情感分析的预测分析的文献中，也主要采用两种路线，传统基于情感词典的方法和基于人工智能模型的方法。传统基于情感词典的方法主要预先准备一个情感词

万方数据

典，然后基于该词典得到一些针对于分析预料的表征，基于这些表征和股票价格数据，再结合分类模型来预测投资者的情绪类别；而基于人工智能模型的方法，主要解决因传统基于构建词典的方法的一些局限性，基于词典的方法构建的字典往往具有专业领域性，移植性较差，同时构建的词典不够完善的话，往往导致提取特征的不准确性，因此提出了基于机器学习的方法，该方法可以避免构建词典的耗时工作，使得使用机器学习的方法自动寻找特征，而且没有领域限制性，移植性较好。在相关的文献中，张世军等人[64]提出了基于网络舆情支持向量机的股票价格预测模型，该方法主要分析新浪微博、腾讯微博和股吧评论，通过关键词词典找到文本中的特征，再通过支持向量机模型获得对股票价格的预测；黄润鹏等人（2015）[65]提出了一个基于微博情绪信息的股票市场预测模型，该模型主要通过对抓取对微博数据进行预处理，再基于情感词典生成情绪时间序列，再结合 svm 模型预测股票走势，相对于不使用情绪序列的模型，准确率有较大提升。另外，彭敏等人（2015）[66]提出了基于情感分析技术的股票研究报告分类模型，该方法主要针对股票报告，采用卡方检验的方法自动选取文本中的特征，再结合分类模型获得预测结果。

2.3 现有研究不足及本文创新之处

前文所述预测模型虽然取得了一定的预测效果，但是石油价格预测准确度仍然不高，这其中的原因是石油价格的影响因素众多且复杂，使得石油价格序列表现出高度复杂性和非平稳性。复杂性主要由于许多全球经济或者国家经济因素的交错影响，这些影响是极其显著而却难以量化的。另一方面非平稳性，主要指石油价格的暴涨暴跌。例如在 1973 年，1978 年和 2008 年，出现过石油价格的暴跌，而在 1986 年和 1998 年，出现过价格的暴涨。石油价格基本是由石油出口国的供给和石油进口国的需求平衡共同决定的，但是往往由政治事件或者经济因素导致的供给的不规律性将会导致石油价格的不规律性。在这种情况下，石油价格的预测将会非常困难，然而石油价格上涨或者下跌的预测对于决策者就显得非常有价值。

而现有的石油价格预测方法往往并没有考虑到除石油供给和需求之外的重要影响因素如，政治和经济因素，而这些影响因素对于石油价格的变化具有非常大的影响，同时如何从众多的新闻数据（或者文本数据）提出有价值的知识以及如何量化这些特征也缺少一个高效准确的方法，因此本文提出了基于情感分析的石油价格走势预测模型，该模型主要基于路透石油相关新闻的分析，通过采用领域关键词词典的方法，得

万方数据

到新闻的情感序列，再通过格兰杰因果检验的方法，得到情感序列和石油价格序列的相关性和滞后期，最后通过机器学习的方法预测石油价格的走势。

因此本文的创新点主要有三处，其一通过基于领域词典的方式捕获与石油相关的在线新闻的新闻情感，来验证新闻情感序列和石油价格的发展变化关系；其二，通过格兰杰因果检验从数量关系上验证新闻情感序列和石油价格的相关关系，并且得到石油价格预测的滞后期；其三，同时引入线性预测模型和非线性人工智能预测模型来验证新闻情感对于石油价格走势的预测能力。

万方数据

第三章 石油价格与新闻情感格兰杰因果分析

3.1 方法概述

3.1.1 情感分析理论概述

在金融市场，信息情感是一种反映投资者和交易者的情感和观点的重要指标。传统金融学的观点是股票价格是上市公司预期利润的贴现，而有效市场假设认为理性投资人会迅速获取撒市场上的消息，并且这些消息会被迅速反应到股票价格上去。但是最近的一些研究成果显示，投资者对于市场消息的反应往往是不理性的，就这导致股票价格通常会过度反应，这就说明了股票价格的影响因素包含投资者情绪这个重要因素。而从行为金融学的角度来讲，股票价格通常是由理性投资者和噪音投资者共同决定的，噪音投资者通常对于资产未来收益率的预期受到投资者情绪的影响，往往这种预期是不准确的，但是理性投资者不能完全消除这种影响，导致股价偏离其基本价值。散户作为封闭式基金的主要交易人，是构成噪音交易者的重要组成部分，因此噪音投资者的乐观或者悲观情绪决定了折价率。因此，从股票价格预测的角度来考虑，充分考虑投资者情绪的预测模型才能更准确的预测股票价格的走势，才能正确的描述市场规律。

情感分析一种度量文本内容情感倾向性的方法。通过情感分析我可以们获得文本内容的情感，但是同时能够知道市场参与者是怎样处理新闻和反馈的。情感分析，也叫做态度挖掘，主要是挖掘出文本内容的正向情感还是负向情感。目前主要有两种情感分析的方法，传统基于词典的方式和基于人工智能模型的方式。

在金融市场领域，情感分析主要应用于股票市场预测。金融文本挖掘方法主要基于词典的方式。这种基于词典的方式通过在给定词典中出现的预先定义的正向词和负向词的频率，在结合一些情感指标的计算规则得到最终到情感分值，然后基于这些情感序列再做相应的分析活着预测。情感指标的计算规则主要有以下几种，Tetlock-Negative 方法，Net-Optimism 方法和本文采用的方法。这里先介绍几个符号表示，$S(A)$ 表示一篇新闻的情感，而 $W_{\text{tot}}(A)$ 表示一篇文章中的所包涵的词的个数，$W_{\text{neg}}(A)$ 表示一篇文章中所包涵的负向词的个数，$W_{\text{pos}}(A)$ 表示一篇文章中的所包涵的正向词的个数。

Tetlock 等人（2008）[50]定义了一种度量投资者情绪的方法，该方法基于 Harvard-IV

万方数据

词典（该词典全部是负向词），计算公式如下所示

$$S_{\mathrm{Tet}}(t)=-\frac{Tet(t)-\mu_{\mathrm{Tet}}}{\sigma_{\mathrm{Tet}}} \quad with \quad Tet(t)=\frac{\sum_A W_{\mathrm{neg}}(A)}{\sum_A W_{\mathrm{tot}}(A)} \qquad \text{式（3-1）}$$

其中，μ_{Tet}是的均值，σ_{Tet}是标准差，是经过训练集所得。而实际上，$Tet(t)$度量了文本中负向词占所有词的比例，并且这个得分是归一化之后的。$S_{\mathrm{Tet}}(t)$表示文本的情绪得分。

Demers 和 Vega(2010)[46]提出了一种不同的计算文本情绪的方法，这种方法主要度量文本中正向词和负向词的差异，正如下面公式所示，

$$S_{\mathrm{NO}}(t)=\frac{\sum_A W_{\mathrm{pos}}(A)-W_{\mathrm{neg}}(A)}{\sum_A W_{\mathrm{tot}}(A)}\in[-1,+1] \qquad \text{式（3-2）}$$

本文采用的情绪计算方法如下，该方法也是 Net-Optimism 方法的一种变体，并且这种方法在股票价格预测领域中使用的非常广泛，其情绪计算得分计算如下，

$$S(t)=\frac{\sum_A W_{\mathrm{pos}}(A)-W_{\mathrm{neg}}(A)}{\sum_A W_{\mathrm{pos}}(A)+W_{\mathrm{neg}}(A)}\in[-1,+1] \qquad \text{式（3-3）}$$

这种计算方法相对于 Net-Optimism 方法，计算时改变了分子，使得文本中正向词和负向词的差异刻画的更加准确。这也是本文的创新点之一，就是通过这种基于领域词典的方法去捕捉石油相关的在线新闻的情感序列，来研究新闻情感和石油价格的相关关系和新闻情感对石油价格的预测能力。

3.1.2 格兰杰因果检验概述

格兰杰因果检验（The Granger causality 是一种基于预测的因果关系的统计概念，是由经济学家克莱夫·格兰杰（Clive W. J. Granger）[70]在 1969 年提出。基于格兰杰因果检验，可以判定一个时间序列是否能辅助预测另一个时间序列。传统的回归方法往往只能揭示是相关关系，但是克莱夫·格兰杰认为通过度量使用一个时间序列的前值预测另外一个时间序列的未来值的能力可以揭示出经济变量之间的因果关系，但是这种因果关系是一种基于“预测”角度的因果关系，并非实际意义的因果关系。它的数学公式是基于随机过程的线性回归得出的。

根据格兰杰因果检验的定义，如果在包含时间序列 X_1 和时间序列 X_2 的历史信息

万方数据

的条件下，引入时间序列 X_1 的历史信息预测序列 X_2 的效果好于仅有序列 X_2 的历史信息预测序列 X_2 的效果，那么就认为序列 X_1 是序列 X_2 的格兰杰原因，即序列 X_1 对于序列 X_2 具有预测能力。

格兰杰因果检验要求两个时间序列同时满足“平稳性”要求，因此对于非平稳时间序列需要对序列进行差分处理，然后再做格兰杰检验，否则可能会出现虚假回归问题。因此通常需要对各指标时间序列的平稳性进行单位根检验(unit root test)，其中增广的迪基—富勒检验(ADF 检验)是一种常用的单位根检验方法。

特别的，两个时间序列 $x(t)$ 和 $y(t)$ 的格兰杰因果关系可以定义如下，同时假设，序列 $y(t)$ 不直接格兰杰引起序列 $x(t)$：

$$\Pr(x(t)\,|\,I(t-1)) = \Pr(x(t)\,|\,I(t-1)-Y_{t-n}^{n})),(t=1,2,...,T), \qquad \text{式（3-4）}$$

这里 $\Pr(x(t)\,|\,I(t-1))$ 表示了当给出二元信息集合 $I(t-1)=\{X_{t-m}^{m},Y_{t-n}^{n}\}$ 的条件下，该集合包括了 m 滞后期的序列 $x(t)$ 和 n 滞后期的序列 $y(t)$ 时，序列 $x(t)$ 的条件概率分布，其中 $X_{t-m}^{m}=\{x(t-m),...,x(t-1)\}$，$Y_{t-n}^{n}=\{y(t-n),...,y(t-1)\}$。另一方面如果公式被假设检验拒绝了，那么可以证明序列 $y(t)$ 过去的信息可以辅助序列 $x(t)$ 未来的估计，也就是说序列 $y(t)$ 能严格引起序列 $x(t)$ 的变化。

在实际中，格兰杰因果检验通过向量自回归模型（VAR）验证：

$$x(t)=a_0+a_1x(t-1)+\cdots+a_mx(t-m)+b_1y(t-1)+\cdots b_ny(t-n)+u(t) \qquad \text{式（3-5）}$$

$$y(t)=a_0{}'+a_1{}'y(t-1)+\cdots+a_n{}'y(t-n)+b_1{}'x(t-1)+\cdots+b_m{}'x(t-m)+v(t) \qquad \text{式（3-6）}$$

这里残差 $u(t)$ 和 $v(t)$ 假设是相互独立的，并且满足均值为 0，标准差为 1 的分布。拿公式来说，一个标准的联合检验（F 检验或者 χ^2 检验）可以通过判断滞后期 $y(t)$ 的相关系数 $b_i(i=1,...,n)$ 是否显著偏离 0，也就是说 $b_i\neq 0,(i\in\{1,...,n\})$，满足则拒绝原假设，即序列 $y(t)$ 能格兰杰引起序列 $x(t)$ 的变化。

因此本文的另外一个创新点就是，通过格兰杰因果检验，可以从数量关系上知道新闻情感和石油价格的变化是否相关，即新闻情感序列和石油价格之间是否具有格兰杰因果关系，新闻情感对于石油价格走势是否具有预测能力，另外，通过最显著的格兰杰因果关系，还可以得到预测的滞后期。

3.2 石油价格数据及新闻数据采集

为了分析新闻情绪和石油价格序列的相关关系，需要石油市场中的新闻。本文中

万方数据

的石油新闻采用的路透社石油专题模块的新闻作为研究语料，由于研究需要，需要大量的新闻数据，因此新闻的获得不能通过人工采集的方式，只能通过爬虫程序实现海量新闻数据的获取。同时，为了后面格兰杰因果检验和预测石油价格走势的需要，还要针对收集到的新闻数据进行预处理。

针对于新闻数据的来源，本文选择了路透社（www.reuters.com），主要有以下几个原因。

其一，路透社新闻有针对于特定商品的新闻板块，例如石油板块，黄金板块和天然气板块，这就使得新闻数据的收集非常方便，同时保证了新闻数据的纯净度，这对以后的分析处理会有很大的帮助。

其二，路透社是一个第三方的新闻出版社，这就保证了新闻的客观性。

其三，相对于报纸新闻，路透社新闻对实时事件反映更迅速，新闻时滞性更小，这对于石油这种商品是非常适合的。

3.2.1 石油数据获取

为了验证新提出的基于情感分析的石油价格走势预测模型的效果，本文选择了西德克萨斯原油期货价格（WTI）作为本文的研究对象，美国西得克萨斯中质原油一直被众多投资者视为国际能源市场的基准价，同时许多国内外的学者也把 WTI 原油价格作为研究对象，主要原因如下，相对于布伦特原油，WTI 原油目前是世界上商品期货中成交量最大的一种，且具有很高的流动性和很高的价格透明度，因此本文选择其作为研究对象。

本文从美国能源署（EIA）获得的从 2008 年 1 月 2 号到 2014 年 12 月 31 号，总共 365 周的 WTI 原油价格数据和相应时期的日度价格，如下图 3-1，图 3-2 所示，从图中本文可以得到几点结论：

第一，从 2008 年到 2014 年 WTI 原油价格的变化表现出了极其大的不规律性和复杂性，这也跟本文在第一章分析的结论一致，石油价格基本是由石油出口国的供给和石油进口国的需求平衡共同决定的，但是往往由政治事件或者经济因素导致的供给的不规律性将会导致石油价格的不规律性。

第二，原油价格的暴涨暴跌变得更加频繁。纵观从 2008 年到现在，不难发现原油价格的暴涨暴跌仿佛已经成了“常态”，这其中的原因除了与石油实际需求有关外，更重要的一个原因来自于石油供应方之间的利益博弈以及地缘政治，而这些对于传统的预测模型来说是难以量化的，因此有效引入这些石

万方数据

油价格的影响因素并量化它们是提高石油价格预测精度的一条重要道路。

第三，WTI 原油周度价格相对于日度价格来说，噪音少，周度走势更加明确，而日度价格影响因素众多，而新闻中情感往往是针对于一些影响石油价格幅度较大的事件的报道和分析，因此使用 WTI 原油周度价格应该更加具有实际意义。

实验中，本文把从 2008 年 1 月 2 号到 2013 年 8 月 3 号的数据（共 292 个样本，80%）作为训练集，剩下时间的数据（共 73 个样本,20%）作为测试集。

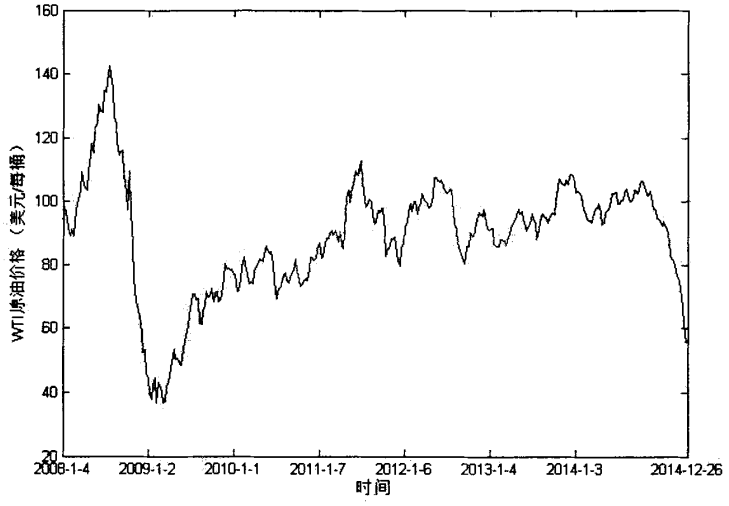

图 3-1.2008 年 1 月 4 号-2014 年 12 月 26 号 WTI 原油周度价格走势

Figure 3-1 WTI weekly futures prices from 4th January 2008 to 26th December 2014

万方数据

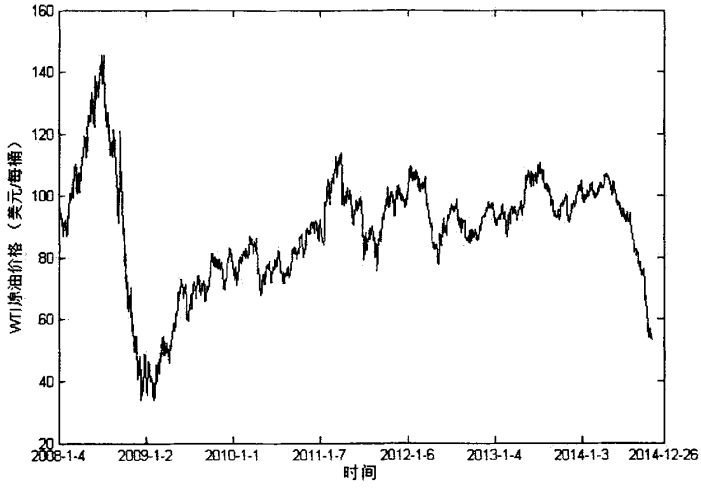

图 3-2. 2008 年 1 月 4 号-2014 年 12 月 26 号 WTI 原油日度价格走势

Figure 3-2 WTI daily futures prices from 4th January 2008 to 26th December 2014

3.2.2 爬虫程序设计

为了更加快捷高效的获取石油新闻数据，本文采用了网页爬虫技术获取海量新闻数据，爬虫程序框架图如下图 3-3 所示。首先，程序的输入主要有两个，一个是带爬的路透社石油专题模块的网址，另外一个是新闻爬取的时间跨度，这个时间跨度主要控制爬取新闻的开始日期和截止日期，同时程序的一些默认参数主要包括请求网页超时时间，最大尝试连接等。其次，爬虫程序会向服务器发请求，根据请求返回状态码判断是否请求成功，如若成功，则首先通过网页分析工具包结合一些规则提出网页中的新闻链接地址，同时把提取到的新闻链接地址加入到待爬队列中去，否则，把该网页链接地址放到失败队列中去，待以后再次尝试。再次，依次获取待爬队列中的网页地址，向服务器发送 http 请求，并把请求结果按照一定的格式（本文中，针对于每篇新闻通过规则提取新闻的日期，标题和内容，并以文本文件的格式进行存储）进行保存。最后，依据截止日期停止爬虫程序。

万方数据

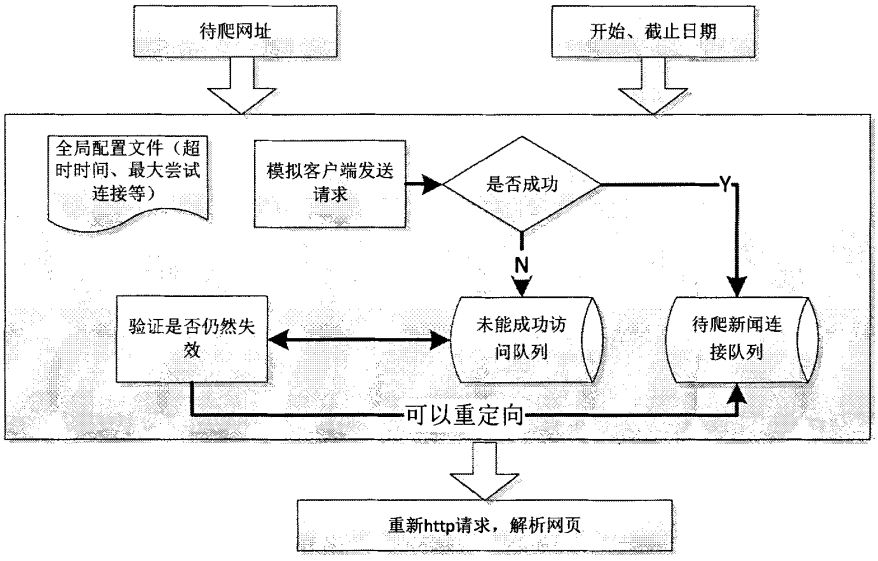

图 3-3. 爬虫程序结构图

Figure 3-3 The Structure diagram of Crawler program

3.3 新闻数据分析

通过爬虫程序，本文最终从路透英国官方网站获得了从 2008 年 1 月 2 号到 2014 年 12 月 31 号（总共 1765 天），总计 163,864 条新闻数据，虽然是石油板块的新闻但是本文发现仍然有很多与石油无关的新闻，例如石油板块中通常也会涉及到天然气等资源的新闻，而这对于分析石油新闻情感是噪音，因此必须对于这些搜集到的新闻必须预处理。为了使得新闻数据尽可能的纯净，本文参考 Wex 等人（2013）的文献自定义了关键词列表（如表 3-1 所示），通过对新闻标题过滤，只保留了高度与石油相关的新闻，总共 14，777 条日度石油新闻，平均每天 8.372 条新闻。

接下来对过滤后的日度新闻，本文合并成周度新闻，并且非工作日的新闻并没有合并到周度新闻中，这主要是因为周末的新闻主要是对本周新闻的重复，然后使用斯坦福分词工具，对合并的新闻语料进行处理，基于文献中的 Henry's 词典，其包括正向词列表（总共 103 个词）和负向词列表（总共 85 个词）提出新闻的特征词，同时考虑到否定句的存在，本文针对于否定词（rather, hardly, couldn't, wasn't, didn't,

万方数据

wouldn't, shouldn't, weren't, don't, doesn't, haven't, hasn't, won't, hadn't, never）做了处理，当有情感词和否定词同时出现时，原情感词的极性翻转，这样本文针对于周度新闻得到了新闻中出现的正负情感词的个数。

表 3-1. 过滤关键词列表

Table 3-1 List of filters

过滤关键词
OILS/CRU/crude oil/HOI/ heating oil/heating diesel/OPEC/oil price

表 3-2. 正向词列表

Table 3-2 List of positive words

positive positives success successes successful succeed succeeds succeeding succeeded accomplish accomplishes accomplishing accomplished accomplishment accomplishments strong strength strengths certain certainly definite solid excellent good leading achieve achieves achieved achieving achievement achievements progress progressing deliver delivers delivered delivering leader leading pleased reward rewards rewarding rewarded opportunity opportunities enjoy enjoys enjoying enjoyed encouraged encouraging up increase increases increasing increased rise rises rising rose risen improve improves improving improved improvements strengthen strengthens strengthening strengthened stronger strongest better best more most above record high higher highest greater greatest larger largest grow grows growing grew grown growth expand expands expanding expanded expansion exceed exceeded exceeding beat beats beating

万方数据

表 3-3. 负向词列表

Table 3-3 List of negative words

negative negatives fail fails failing failure weak weakness weaknesses difficult difficulty hurdle hurdles obstacle obstacles slump slumps slumping slumped uncertain uncertainty unsettled unfavorable downturn depressed disappoint disappoints disappointing disappointed disappointment risk risks risky threat threats penalty penalties down decrease decreases decreasing decreased decline declines declining declined fall falls falling fell fallen drop drops dropping dropped deteriorate deteriorates deteriorating deteriorated worsen worsens worsening weaken weakens weakening weakened worse worst low lower lowest less least smaller smallest shrink shrinks shrinking shrunk below under challenge challenges challenging challenged

3.4 石油走势与新闻情感格兰杰因果关系分析

情感分析和 WTI 价格序列归一化之后，本文采用 HP 滤波法去掉这两个序列的噪音，保留趋势，然后把去掉噪音后的两个序列画在了图 3-4 中。从图中可以得到三个主要的结论。第一，两个序列协同发展，其正相关系数为 0.4041，较高的相关性表明新闻情感的变化同石油价格走势的变化具有很大的相关度，这表明新闻情感相对于石油价格来说有重要的参照作用。第二，新闻情感的变化早于石油价格的变化，但是这不局限于一个特定的时间。例如在 2008 年时，新闻情感的变化大约早于石油价格变化三到四周，这个原因跟 2008 年世界经济危机有着重要关系，世界经济形势的下滑导致新闻情绪的下降，新闻情感的下降导致石油需求的疲软，在这种负面的环境下，两者互相影响，而且影响速度是递增的；而从 2009 年之后，本文发现这个时间明显变长大约变为六到八周，到 2014 年时又大约回归了三到四周，但是这个时间的变化并没有表现出一个特别明显的规律，但是无论怎样，可以看到，新闻情感对于石油价格具有预测能力，也就是说新闻情感的变化能有解释石油价格的变化。第三，新闻情

万方数据

感的正负和石油价格的上涨和下跌有对应关系，同时新闻情感的变化值也具有这种关系，也就是说，新闻情感值的正负和新闻情感值变化的正负同时影响石油价格的变化。

例如 2008 年石油价格在上涨到最高值之后，开始下降，但是本文发现这个时候新闻情感的值其实是大于零的，但是新闻情感在不早之前已经开始下降，新闻情感的分值代表了投资者对于石油走势的态度，虽然投资者依然对当前的石油走势保持了乐观态度，但是这种乐观态度的递减实际是表示了投资者对于石油价格未来的走势越来越悲观，因此石油价格在表现也是逐渐下降，从 2009 年初的时候也可以看到类似的情况，虽然新闻情感值为负值，但是情感负值是逐渐增大的，也就说虽然市场投资者依然对石油价格走势保持悲观态度，但是这种悲观程度是逐渐减弱的，表现在石油价格上就是石油价格的不断上涨。虽然从图中可以分析到新闻情感对于石油价格的走势预测具有解释能力和预测能力，但是本文为了验证新闻情感的变化是否真正与石油价格变化相关，本文采用了格兰杰因果检验。

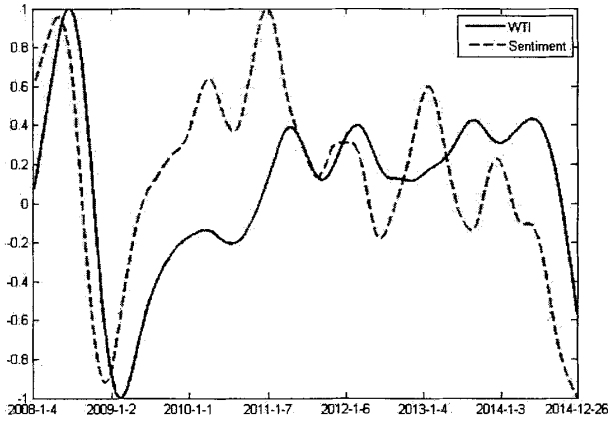

图 3-4.2008.01.04-2014.12.26 WTI 归一化周度价格和新闻情感的变化图

Figure 3-4WTI weekly futures prices after normalization and sentimental series from 4th January 2008 to 26th December 2014.

在格兰杰因果检验之前，鉴于格兰杰因果检验的先决条件是时间序列满足平稳性要求，为此进行了 ADF 检验，来检验这两个序列的平稳性，结果见表 3-4. 结果表明两个原始序列在相同的 1%的显著水平下并不满足平稳性要求，为此验证两个序列在一阶差分的情况下，结果表明在 1%的显著水平下都是平稳的。随后，进行了协整检验，

万方数据

来验证这两个序列是否具有协整关系，结果见表 3-5。同样结果表明，在 1%的置信水

表 3-4.单位根检验结果

Table 3-4 Unit root tests on individual series

t-Stat. (*p*-value)	Series in levels	Series in first difference
News sentiment	-2.9364 (0.0422)	-3.6584 (**0.0051**)
WTI crude oil price	-2.2317(0.1955)	-3.7247 (**0.0041**)

表 3-5. 协整检验结果

Table 3-5 Unit root tests on residual series

	Series in levels	Series in first difference
t-Stat. (*p*-value)	-2.7899 (0.0607)	-3.7182 (**0.0042**)

表 3-6. 新闻情感和 WTI 序列的格兰杰因果检验结果

Table 3-6 Granger-causality correlation between sentimental series and WTI

	Lags								
	1	2	3	4	5	6	7	8	9
χ^2- Stat.	297.2693	24.7576	56.4907	3.4391	3.2588	3.4608	4.9915	5.8369	9.7300
p-value	**0.0000**	**0.0000**	**0.0000**	0.4872	0.6602	0.7492	0.6610	0.6655	0.3728

平下，在一阶差分情况下两个序列具有协整关系。最后，对于这两个序列进行了格兰杰因果检验，结果见表 3-6。可以看到，新闻情感和石油价格在滞后期为 1, 2 和 3 的

万方数据

时候（p-value<0.01），具有明显的格兰杰因果关系，而在其他滞后期的时候，没有显著的格兰杰因果关系。因此，这也验证了之前在前面一小节内得到的结论，即新闻情感的变化同石油价格走势的变化具有很大的相关度，这表明新闻情感相对于石油价格来说有重要的参照作用，新闻情感的变化早于石油价格的变化，并且在短时间（3-4周）内新闻情感对石油价格具有预测能力。同时，在后面的基于新闻情感和人工智能模型预测石油价格走势预测时，本文基于 AIC 最小原则选择 3 作为滞后期。

3.5 本章小结

本章首先分析了情感分析的基本理论，指出在金融市场，信息情感是一种反映投资者和交易者的情感和观点的重要指标，充分考虑投资者情绪的预测模型才能更准确的预测股票价格的走势，才能正确的描述市场规律。同时指出，情感分析一种度量文本内容情感倾向性的方法。的确，通过情感分析获得文本内容的情感，但是同时能够知道市场参与者是怎样处理新闻和反馈的。情感分析，也叫做态度挖掘，主要是挖掘出文本内容的正向情感还是负向情感。目前主要有两种情感分析的方法，传统基于词典的方式和基于人工智能模型的方式。在金融市场领域，情感分析主要应用于股票市场预测。金融文本挖掘方法主要基于词典的方式。这种基于词典的方式通过在给定词典中出现的预先定义的正向词和负向词的频率，在结合一些情感指标的计算规则得到最终到情感分值，然后基于这些情感序列再做相应的分析活着预测。情感指标的计算规则主要有以下几种，Tetlock-Negative 方法，Net-Optimism 方法和本文采用的方法。其次，仔细介绍了格兰杰因果检验的定义、理论基础和检验步骤，同时指出时间序列在进行格兰杰因果检验必须满足“平稳性”条件，并结合本文的研究内容进行了分析。

为了分析新闻情绪和石油价格序列的相关关系，需要石油市场中的新闻。本文中的石油新闻采用的路透社石油专题模块的新闻作为研究语料，由于研究需要，需要大量的新闻数据，因此新闻的获得不能通过人工采集的方式，只能通过爬虫程序实现海量新闻数据的获取。同时，为了验证新提出的基于情感分析的石油价格走势预测模型的效果，本文选择了西德克萨斯原油期货价格（WTI）作为本文的研究对象，通过对获得的从 2008 年 1 月 2 号到 2014 年 12 月 31 号，总共 365 周的 WTI 原油价格数据的分析，可以知道第一，从 2008 年到 2014 年 WTI 原油价格的变化表现出了极其大的不规律性和复杂性，并且原油价格的暴涨暴跌变得更加频繁，同时考虑到日度数据

万方数据

噪音较多，本文实验数据选择周度石油价格作为研究对象。

然后对于爬虫程序获得新闻数据，本文首先采用关键词过滤的方式过滤掉与原油不相关的新闻，保证后面的新闻情感提取的有效性，然后对于获得日度新闻进行合并获得周度新闻，然后基于情感词典和一些规则提出新闻中的情感特征。

通过对获取的新闻的预处理，再根据本文采用的情感计算的方法，获得了从2008年1月2号到2014年12月31号的石油新闻的情感序列，通过与石油价格序列的分析对比，得到了三个主要的结论，即：

（1）新闻情感和石油价格大体上存在着相对较高的，正相关相关关系，说明新闻情感是石油价格变化的一个参照。

（2）新闻情感的变化早于石油价格变化，但是这个时间并不固定，说明新闻情感对于石油价格具有预测能力。

（3）新闻情感的正负和石油价格的上涨和下跌有对应关系，同时新闻情感的变化值也具有这种关系，也就是说，新闻情感值的正负和新闻情感值变化的正负同时影响石油价格的变化。

接下来，为了验证新闻情感序列的变化能够引起石油价格的变化，本文使用了格兰杰因果检验，又因为格兰杰因果检验要求时间序列具有平稳性，因此又分别进行了单位根检验和协整检验，格兰杰因果检验结果表明，在短时间内，新闻情感对于石油价格的变化具有重要的预测能力。

万方数据

万方数据

第四章 基于情感分析的石油价格走势预测

前面一章，通过爬虫程序，本文搜集到了海量的石油新闻数据，在经过第三章的石油价格与新闻情感格兰杰因果检验之后，可以推断新闻情感和石油价格存在着格兰杰因果关系，特别地新闻情感的变化早于石油价格的变化，即新闻情感对于石油价格的变化具有预测能力，但是格兰杰因果检验只能断定新闻情感序列和石油价格序列的线性关系，而众所周知石油价格因素错综复杂，因此石油价格更倾向表现出一种非线性，为了捕捉新闻情感对于石油价格的线性和非线性影响，本文同时引入了线性（LogR）和非线性预测模型（DT,BPNN 和 SVM），把获得的情感序列加入到预测石油价格走势的模型当中。

4.1 预测模型概述

前面的格兰杰因果分析可能表明新闻情感对于石油价格的变化具有一定的预测能力。然而，前面的分析也只是说明新闻情感和石油价格之间具有线性关系，然而知道新闻情感和石油价格之间更可能存在的是一种非线性的关系，为了更好的检验这种非线性关系和新闻情感对于石油价格是否具有预测能力，本文选择了一些优秀机器学习模型（支持向量机，决策树，逻辑回归和神经网络）来基于新闻情感预测石油价格变化，这些模型能够很好的处理非线性数据，并且已经广泛应用在股票预测中。接下来，将会仔细介绍每一个分类模型。

逻辑回归模型是有 David Cox（1958）[71]提出的，也是现在商业界比较流行的机器学习方法，逻辑回归是一种监督学习方法。二元逻辑回归模型基于一个或多个变量决定一个事件发生的可能性，注意这里并不是数学上的概率，不过可以把它当作事件发生的概率。逻辑回归使用逻辑函数度量一个类别变量和一个或多个独立变量的关系，该函数满足累积逻辑概率分布。跟线性回归不同的是，逻辑回归是一个非线性函数，其拟合能力就更强大了，并且它是连续函数，可以求导，求导意味着很快就可以找到极值点，因此逻辑回归的重要思想是利用求导得到梯度，然后利用梯度下降法更新参数。从下图的可以看出，逻辑回归比较适合做二分类，因为它严重两极分化。从公式中可以看出，逻辑回归本质上讲也是一种线性回归的方法，即先把特征线性求和，然后通过非线性函数映射，最后使用逻辑函数来预测，该函数将连续值映射到 0 到 1 上。

逻辑回归到基本原理如下，首先，寻找合适的预测函数（分类函数），表示为 h

万方数据

函数，这一般跟研究的内容有很大关系，比如在本文中研究的是石油价格走势的预测。其次，构造一个损失函数。用该函数表示预测的输出和训练数据类别的误差，可以是以差值的形式，也可以是其他形式。然后把整个训练集的损失作为误差。最后，根据梯度下降算法计算更新模型参数。

具体过程如下，

首先，预测函数使用了 Logistic 函数，也叫做 Sigmoid 函数，函数形式为：

$$g(z)=\frac{1}{1+e^{-z}} \qquad \text{式（4-1）}$$

其对应的函数图像是一个取值在 0 和 1 之间的 S 曲线。

接下来需要确定数据划分的的边界，针对于线性边界，形式如下，

$$\theta_0+\theta_1x_1+\cdots+\theta_nx_n=\sum_{i=0}^{n}\theta_ix_i=\theta^Tx \qquad \text{式（4-2）}$$

构造预测函数为：

$$h_\theta(x)=g(\theta^Tx)=\frac{1}{1+e^{-\theta^Tx}} \qquad \text{式（4-3）}$$

函数的值表示预测类别为 1 的概率，因此对于输入 x 预测的分类结果分别为 1 和 0 的概率分别为：

$$\begin{aligned}P(y=1\mid x;\theta)&=h_\theta(x)\\P(y=0\mid x;\theta)&=1-h_\theta(x)\end{aligned} \qquad \text{式（4-4）}$$

上面两个公式综合起来得到下面的公式，

$$P(y\mid x;\theta)=(h_\theta(x))^y(1-h_\theta(x))^{1-y} \qquad \text{式（4-5）}$$

取似然函数为：

$$\begin{aligned}L(\theta)&=\prod_{i=1}^{m}P(y(i)\mid x(i);\theta)\\&=\prod_{i=1}^{m}(h_\theta(x^{(i)}))^{y^{(i)}}(1-h_\theta(x^{(i)}))^{1-y^{(i)}}\end{aligned} \qquad \text{式（4-6）}$$

对数似然函数为：

$$\begin{aligned}l(\theta)&=\log L(\theta)\\&=\sum_{i=1}^{m}(y^{(i)}\log h_\theta(x^{(i)}))+(1-y^{(i)})\log(1-h_\theta(x^{(i)}))\end{aligned} \qquad \text{式（4-7）}$$

万方数据

最大似然估计就是要求使得$l(\theta)$取得最大值时的θ，这里再做一下变换，如下所示，

$$J(\theta)=-\frac{1}{m}l(\theta) \qquad \text{式（4-8）}$$

所以$J(\theta)$取最小值时的θ为要求的最佳参数。

接下来，使用梯度下降法求$J(\theta)$的最小值

决策树是一个非常流行的分类算法，并且已经广泛地应用到现实世界中[72]。决策树算法同样是一个监督学习算法，并且是一种以实例为基础的学习分类规则的算法。最终它会把从训练数据中学习出的分类函数表示成一棵树的形式。常见的决策树算法有 ID3 算法，C4.5 算法和 CART 算法。

ID3 算法是有 Quinlan[73]在 1986 年提出的一种基于信息熵的决策树学习算法，它的前身是 CLS 算法。Quinlan 在决策树算法中引入了香农的信息论的概念。ID3 算法的核心观点是利用信息增益作为属性选择的标准，每次选择信息增益最大的属性作为决策树的节点，并根据这个节点的不同属性值构建树的分支，然后递归的构建完整个树。

信息增益计算如下：

假定属性 A 可以有v个不同的取值$\{a_1,a_2,\ldots,a_v\}$，并基于属性 A 把S 分解成子集$\{S_1,S_2,\ldots,S_v\}$。假定把属性 A 作为测试属性，那么子集就是从上面S 得到的子集。假定s_{ij}表示子集S_j样本属于类别C_i的样本数。那么基于属性 A 得到的信息熵定义如下：

$$E(A)=-\sum_{j=1}^{v}\frac{s_{1j}+s_{2j}+\ldots s_{mj}}{s}I(s_{1j},s_{2j},\ldots,s_{mj}) \qquad \text{式（4-9）}$$

这里$\frac{s_{1j}+s_{2j}+\ldots+s_{mj}}{s}$表示子集 j 的权重，它等于子集 j 中样本数（这些的样本的 A 属性值为a_j）除以总的样本数。该信息熵越小，划分的子集越纯净。对于子集S_j，它的期望信息通过下式定义：

$$I(s_{1j},s_{2j},\ldots s_{mj})=-\sum_{i=1}^{m}p_{ij}\log p_{ij} \qquad \text{式（4-10）}$$

这里$p_{ij}=\frac{s_{ij}}{|s_j|}$表示样本$S_j$属于类别$C_i$的概率。然后通过该子集的信息熵和期望信息得到信息增益，表示如式（4-11）所示：

$$Gain(A)=I(s_1,s_2,\ldots,s_m)-E(A) \qquad \text{式（4-11）}$$

万方数据

计算每一个属性的信息增益，然后选择信息增益最大的属性作为测试属性。ID3算法具有算法简单、理论清晰、强学习能力和易于构建的特点。缺点主要有：（1）属性选择偏爱问题，它会倾向于选择属性值多的属性，这对于分类问题是没有意义的。（2）对噪音非常敏感，当训练集数量增加或减少时，学习到的决策树的结构经常会变。（3）在属性选择时，没有考虑到属性之间的相关关系。（4）对于学习逻辑回归函数的能力欠缺。

BP 神经网络算法是有 Rumelhart 等人[74]（1986）提出的一种基于梯度下降的学习网络权重的多层感知的前馈神经网络。刚开始训练时，网络的权重随机初始化。然后基于训练数据的网络输出与期望输出之间的误差不断的调整网络权重，优化网络。在具体迭代的过程中，输入数据输入到网络结构中，经过前馈过程，得到一个输出模式，然后根据这个输出结果和期望输出的结果的误差通过后馈过程传递误差，以至于更新网络权重。这个学习过程一直持续到整个训练集的输出结果同期望输出结果的均方根误差小于预先设定的值或者达到最大训练次数。Rumelhart 的方法如下面公式所示：

$$o_{pj}=\frac{1}{1+\exp\{-(\sum_{i} w_{ji}o_{pi}+\theta_j)\}}, \qquad \text{式（4-12）}$$

这里当一个输入模式 p 输入到神经网络时，每一个神经元节点的输出使用上面的 sigmod 激活函数得到。这里，o_{pj} 表示神经元 j 的激活函数，p 表示该神经元的输入，w_{ji} 代表从神经元 i 当神经元 j 的链接权重，θ_j 代表神经元 j 的偏置。然后通过反向传播更新误差，公式如下：

$$\Delta w_{ji}(n+1)=\eta\cdot\delta_{pj}\cdot o_{pi}+\alpha\cdot\Delta w_{ji}(n), \qquad \text{式（4-13）}$$

这里 n 代表反向传播次数，η 表示学习速率，δ_{pj} 表示神经元 j 的误差，α 代表了动量系数。这里的输出层神经元的误差 δ_{pj} 是通过计算神经元 j 的目标输出同实际输出的差值得到的，公式如下：

$$\delta_{pj}=(t_{pj}-o_{pj})\cdot o_{pj}\cdot(1-o_{pj}). \qquad \text{式（4-14）}$$

隐含层神经元 j 的误差是通过计算下一隐含层的神经元和该神经元的误差，公式如下：

$$\delta_{pj}=o_{pj}\cdot(1-o_{pj})\cdot\sum_{k}\delta_{pk}w_{kj}. \qquad \text{式（4-15）}$$

实际上，该算法在运用时是这样运行的：每一个输入样本输入到网络中，输出层得到实际输出，然后通过反向传播更新全部网络的参数，然后下一个样本输出，重复同样的过程。参数η 和α，是用户可以自己设定的参数。

万方数据

支持向量机（Support Vector Machines, SVM），另外一种人工智能算法，它是由Vapnik(1995)提出的基于结构风险最小化原则的算法，即最小化误差最大限值，具有很好的泛化能力及卓越的分析能力，特别是针对小样本数据。鉴于出色的学习能力和稳健的统计理论基础，支持向量已经被广泛应用到预测任务中去，甚至是处理复杂的数据样本。它的基本观念是基于非线性映射函数把原始数据映射到高维空间，然后在此特征空间找到间隔最大的超平面（分类平面或者分类函数），使得该平面附近的点间的距离更大，这使得支持向量机有别于感知机；上面提到的间隔最大策略，最终可转化为凸二次规划(convex quadratic programming)求解问题。

假设训练数据有 n 个观测样本，也就是{(x_1, y_1), …, (x_n, y_n)}，其中 x_i (i=1, …, n)表示输入，y_i (i=1, …, n)表示输出，那么支持向量机分类问题可以表述为如下所示：

$$\begin{cases} \min \quad J(\boldsymbol{a},\boldsymbol{b},\xi)=(1/2)\boldsymbol{a}^T\boldsymbol{a}+\gamma\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad y_i\left[\varphi(x_i)\cdot a_i+\boldsymbol{b}\right]\geq 1-\xi_i, \\ \qquad \xi_i\geq 0, i=1,2,...,n, \end{cases} \qquad \text{式（4-16）}$$

这里 $\boldsymbol{a}=\{a_1,...,a_n\}$ 表示超平面向量，b 表示偏置，$\varphi(x_i)$ 代表把输入数据转化到高维空间的非线性映射函数。$\boldsymbol{\xi}=\{\xi_1,\xi_2,...,\xi_n\}$ 表示可容忍的最小的分类误差，γ 表示控制最大间隔和可容忍的最小分类误差的正则化因子。

在实际中，多种多样的核函数可以作为非线性映射函数来简化映射过程。其中最常用的核函数就是 RBF 核函数，公式如下：

$$K(x_i,x_j)=\exp(-\|x_i-x_j\|/2\sigma^2) \qquad \text{式（4-17）}$$

其中参数 σ^2 是需要人为设定的。在本文中，使用了该核函数。

因此，在支持向量机模型中，总共有两个参数需要仔细测验已达到好的预测效果，分别是参数 γ 和 RBF 核函数的参数 σ^2。网格选参是一个不错的寻参办法，因此本文使用网格法。

在没有其他额外的变量下，基于历史石油价格序列 $X_t=\{x_t,x_{t-1},...,x_{t-(m-1)}\}$ 预测步长为 h 的情况下的石油价格走势 $\hat{y}_{t+h}$ 的公式可以表述为如下：

万方数据

$$\hat{y}_{t+h} = f(X_t) = f(x_t, x_{t-1}, ..., x_{t-(m-1)}), \quad \text{式（4-18）}$$

其中 $\hat{y}_t$ 表示在 t 时刻的石油价格走势预测值，m 表示滞后期，h 是步长。特别的，$\hat{y}_t = 0$ 预示着石油价格将要下降（$x_t < x_{t-1}$），相反 $\hat{y}_t = 1$ 预示着石油价格将要上升（$x_t > x_{t-1}$）。引入新闻情感 s_t 和预测滞后期 l，则预测模型可以从式（4-18）推导出来如下：

$$\hat{y}_{t+h} = f\{S_t, X_t\} = f(s_{t-l+1}, ..., s_t, x_{t-m+1}, ..., x_t). \quad \text{式（4-19）}$$

因此，本文的最后一个创新之处在于同时引入线性预测模型和非线性预测模型来研究新闻情感对于石油价格走势的预测能力，在对未来石油价格走势预测时，不只考虑历史价格序列，同时引入历史新闻情绪序列辅助石油价格走势的预测。

4.2 实验设计

本文选取了分类准确率（*PCC*）作为模型好坏的评价指标，

$$PCC = \frac{\sum_{t=1}^{M} a_t}{M}, \; at = \begin{cases} 1, \hat{y}_t = y_t \\ 0, \hat{y}_t \neq y_t \end{cases}, \quad \text{式（4-20）}$$

其中 M 表示测试集的数量，$\hat{y}_t = \{0,1\}$ 和 $y_t = \{0,1\}$ 分别表示在 t 时刻石油价格走势的预测值和真实值，特别的，$\hat{y}_t = 1$（或者 $\hat{y}_t = 0$）表示预测石油价格将会上升（或者下降），y_t 同理。从统计检验方面考虑，二项分布检验用来检验模型的 *PCC* 是否显著高于基准 *PCC'*，这里设定的基准 *PCC'* 等于 50%。检验公式如下：

$$z = \frac{PCC - PCC'}{\sqrt{\frac{(1-PCC)\cdot PCC'}{M}}}. \quad \text{式（4-21）}$$

为了研究新闻情感对于石油价格走势的预测能力，分别引入总共两类预测模型来进行实验研究，分别是基准模型（标记为 Type-Ⅰ）和本文新提出的预测模型（标记为 Type-Ⅱ）。对于基准模型，如式（4-18）所示，其输入只包括历史石油价格序列，而对于本文新提出的模型，如式（4-19）所示，新闻情感被当做一个重要的预测变量引入到石油价格走势预测模型中。因此，考虑到上文提到的分类模型（LogR、DT、BPNN 和 SVM），总共有八个预测模型，分别是 LogR-Ⅰ，DT-Ⅰ，BPNN-Ⅰ和 SVM-

万方数据

Ⅰ作为基准模型，而 LogR-Ⅱ，DT-Ⅱ,BPNN-Ⅱ和 SVM-Ⅱ作为新提出的预测模型。

关于模型的参数设置问题，针对于 Type-Ⅰ和 Type-Ⅱ模型，所有的分类模型采用的相同的参数设定标准。对于决策树算法，本文采用了 ID3 算法。对于神经网络算法，隐含神经元节点的个数采用$2*n+1$的方式设定，其中 n 表示输入变量的个数。对于支持向量机模型，采用 RBF 函数作为核函数，并且使用网格寻参的方法寻找合适参数γ和σ^2。

4.3 预测结果分析

表 4-1. 各个预测模型结果对比

Table 4-1 Prediction performances of several popular classification models

	BPNN		**LogR**		**DT**		**SVM**	
	PCC	***p*-value**	***PCC***	***p*-value**	***PCC***	***p*-value**	***PCC***	***p*-value**
Type-Ⅰ	0.5644	0.1600	0.6301	0.0340	0.5479	0.4830	0.5753	0.2420
Type-Ⅱ	0.5986	**0.0020**	0.6712	**0.0050**	0.6575	**0.0100**	0.6712	**0.0050**

八个预测模型的预测结果在表 4-1 中。结果表明，所有的 Type-Ⅱ模型的预测结果优于 Type-Ⅰ模型，也就是说新闻情感对于石油价格变化的预测具有重要的影响，即所有加入新闻情感的模型预测准确率要明显高于只使用历史石油价格预测的模型。特别地，所有 Type-Ⅱ模型的预测表现显著地优于相对应的 Type-Ⅰ模型。Type-Ⅱ模型（LogR-Ⅱ，DT-Ⅱ,BPNN-Ⅱ和 SVM-Ⅱ）的平均分类准确率为 64.9625%，远高于基准模型（LogR-Ⅰ，DT-Ⅰ，BPNN-Ⅰ和 SVM-Ⅰ）的平均分类准确率 57.9425%。

同时，在所有的基础模型，即没有加入新闻情感的预测模型中，逻辑回归模型准确率最高为 63.01%，其次是使用网格选参的支持向量机模型为 57.53%，再其次是神经网络的 56.44%，最后是决策树的 54.79%，结果表明在基准模型中，线性预测模型预测效果最好，非线性模型引入并没有提高石油价格走势预测的准确率。而在引入新闻情感辅助石油价格走势预测中后，每一个分类模型都有了准确率的提升，其中决策树模型提升的幅度最大，达 10.96%，神经网络模型准确率提升了 3.42%，逻辑回归模

万方数据

型提升了 4.11%，基于网格选参的支持向量机模型提升了 9.59%，结果表明无论对于线性预测模型还是非线性预测模型，新闻情感的引入都提高了预测准确率，而且从提升幅度来讲，非线性模型效果提升幅度远高于线性模型的提升幅度。

4.4 预测结果统计检验分析

此外，二项分布统计检验结果表明，引入新闻情感的模型预测效果显著优于基准模型。特别的，所有 Type-Ⅱ模型，在 1%的显著水拼下，p 值显著低于相对应的 Type-Ⅰ模型的 p 值，这说明所有 Type-Ⅱ模型（LogR-Ⅱ，DT-Ⅱ,BPNN-Ⅱ和 SVM-Ⅱ）针对于石油价格走势预测获得了更好的预测结果，而对于 Type-Ⅰ模型（LogR-Ⅰ，DT-Ⅰ，BPNN-Ⅰ和 SVM-Ⅰ）来讲，在 1%的显著水平下，所有模型的 p 值都没有满足要求，这也说明所有基准模型的预测效果并没有通过统计检验，即只采用历史石油价格序列并不能有效预测石油价格的走势变化，而引入新闻情感可以显著提高石油价格走势的预测准确率。

4.5 本章小结

本章首先介绍了本文所使用的分类模型的基本原理和求解方法，分别介绍了逻辑回归模型、决策树模型、神经网络模型和支持向量机模型，为下文的石油价格走势的预测做出铺垫。为了捕捉新闻情感对于石油价格的线性和非线性影响，把获得的情感序列加入到预测石油价格走势的模型当中。根据格兰杰因果检验得到的滞后期，针对有所有分类模型，分别构建了两类预测模型，基准模型（Type-Ⅰ）和引入新闻情感的预测模型（Type-Ⅱ），并选择分类准确率作为模型评价指标，并对预测结果进行了假设统计检验。结果表明，所有的 Type-Ⅱ模型的预测结果优于 Type-Ⅰ模型，也就是说新闻情感对于石油价格变化的预测具有重要的影响，即所有加入新闻情感的模型预测准确率要明显高于只使用历史石油价格预测的模型。结果同时表明在基准模型中，线性预测模型预测效果最好，非线性模型引入并没有提高石油价格走势预测的准确率。无论对于线性预测模型还是非线性预测模型，新闻情感的引入都提高了预测准确率，而且从提升幅度来讲，非线性模型效果提升幅度远高于线性模型的提升幅度。此外，二项分布统计检验结果表明，引入新闻情感的模型预测效果显著优于基准模型。

万方数据

第五章 结论与展望

5.1 主要结论

鉴于石油已经成为世界上最重要的商品之一，石油价格的变化对于世界经济的发展具有重要影响。无论对于石油出口国还是输出国，都极大地促进了当地的经济的发展。

因此，无论是国内学者还是国外学者都积极投入到石油价格预测研究中。伴随着石油价格的不稳定变化和世界经济的放缓，从石油价格变化行为理论研究，到石油价格和经济指标相关性分析理论研究，再到石油价格预测理论研究，国内外学者涌现了一大批优秀的研究成果。其中，石油价格变化行为理论，旨在研究石油价格的影响因素以及各因素的相互作用，这对于弄清石油价格变化的机理具有重要意义。而石油价格和经济指标相关性分析主要研究石油价格同经济发展指标的相关关系，这些理论对于研究石油价格对于世界经济的影响具有重要作用。对于石油价格预测领域来说，旨在探索一系列的石油价格预测方法的研究。这些预测模型虽然取得了一定的预测效果，但是石油价格预测准确度仍然不高，这其中的原因是石油价格的影响因素众多且复杂，使得石油价格序列表现出高度复杂性和非平稳性。复杂性主要由于许多全球经济或者国家经济因素的交错影响，这些影响是极其显著而却难以量化的。另一方面，石油价格的暴涨暴跌时常发生。例如在 1973 年，1978 年和 2008 年，出现过石油价格的暴跌，而在 1986 年和 1998 年，出现过价格的暴涨。石油价格基本是由石油出口国的供给和石油进口国的需求平衡共同决定的，但是往往由政治事件或者经济因素导致的供给的不规律性将会导致石油价格的不规律性。在这种情况下，石油价格的预测将会非常困难，然而石油价格上涨或者下跌的预测对于决策者就显得非常有价值。

同时，随着互联网和大数据技术的发展，基于大数据的市场预测已经越来越受到重视。如何有效的管理和使用这些大规模的数据已然成为了一个重要的研究课题。近些年来，学术界，工业界和政府已经开始关注并使用大数据解决传统问题。特别的对于金融市场，一大批研究学者已经基于大数据的处理来研究金融市场的运作规律。根据有效市场的假设，金融信息对与金融市场的波动性具有重要影响。事实上，新闻内容代表了对金融市场未来趋势的实时评估。新闻的内容将会影响金融投资者的投资行为，进而影响金融市场。石油，作为一种特殊的商品，其影响因素，例如经济，军事和政治因素，自然灾害，投机和供给需求，都汇集在互联网上。所有的这些著名网站，

万方数据

产生大量数据。如果能好好利用这些数据，那么对于石油价格趋势的预测将产生重要的影响。

新闻的文本内容提供除了关于事件事实的报道，还包括语言的语调或者情绪。而这些语调或者情绪反映了市场对石油价格走势的理解和判断，而这些新闻通常来源于市场重要的专家或者学者，因此这些语调代表了市场参与人员对石油市场的重要的态度，因此这些情绪对于石油价格走势的判断具有重要作用。因此本文考虑把新闻情绪引入到石油价格走势的预测中。

在这种背景下，本文提出了基于情感分析的国际原油价格走势预测模型，该模型主要基于石油相关新闻的分析，通过采用领域关键词词典的方法，得到新闻的情感序列，再通过格兰杰因果检验的方法，得到情感序列和石油价格序列的相关性和滞后期，最后通过机器学习的方法预测石油价格的走势。

大体上来说，新提出的预测方法主要包括三步：情感分析和归一化，格兰杰因果检验和石油变化趋势预测。第一步，基于情感分析方法获得新闻文本的情感倾向性，同时对于石油价格的序列进行归一化处理。第二步，使用格兰杰因果检验验证新闻情感的变化是否决定石油价格的变化，同时确定滞后期。第三步，使用一些优秀的机器学习模型，例如支持向量机，决策树，逻辑回归和神经网络算法结合新闻情感预测石油价格变化的趋势。

为了验证新提出的基于情感分析的石油价格走势预测模型的效果，本文选择了西德克萨斯原油期货价格（WTI）和路透社石油模块大新闻作为本文的研究对象，通过对获取的新闻的预处理，再根据本文采用的情感计算的方法，获得了从 2008 年 1 月 2 号到 2014 年 12 月 31 号的石油新闻的情感序列，通过与石油价格序列的分析对比，本文得到了三个主要的结论，即：

（1）新闻情感和石油价格大体上存在着相对较高的，正相关相关关系，说明新闻情感是石油价格变化的一个参照。

（2）新闻情感的变化早于石油价格变化，但是这个时间并不固定，说明新闻情感对于石油价格具有预测能力。

（3）新闻情感的正负和石油价格的上涨和下跌有对应关系，同时新闻情感的变化值也具有这种关系，也就是说，新闻情感值的正负和新闻情感值变化的正负同时影响石油价格的变化。

接下来，为了验证新闻情感序列的变化能够引起石油价格的变化，使用了格兰杰因果检验，又因为格兰杰因果检验要求时间序列具有平稳性，因此又分别进行了单位根检验和协整检验，格兰杰因果检验结果表明，在短时间内，新闻情感对于石油价格

万方数据

的变化具有重要的的预测能力。

最后为了捕捉新闻情感对于石油价格的线性和非线性影响，同时引入了线性（LogR）和非线性预测模型（DT,BPNN 和 SVM），把获得的情感序列加入到预测石油价格走势的模型当中。根据格兰杰因果检验得到的滞后期，针对有所有分类模型，分别构建了两类预测模型，基准模型（Type-Ⅰ）和引入新闻情感的预测模型（Type-Ⅱ），并选择分类准确率作为模型评价指标，并对预测结果进行了假设统计检验。结果表明，所有的 Type-Ⅱ模型的预测结果优于 Type-Ⅰ模型，也就是说新闻情感对于石油价格变化的预测具有重要的影响，即所有加入新闻情感的模型预测准确率要明显高于只使用历史石油价格预测的模型。结果同时表明在基准模型中，线性预测模型预测效果最好，非线性模型引入并没有提高石油价格走势预测的准确率。无论对于线性预测模型还是非线性预测模型，新闻情感的引入都提高了预测准确率，而且从提升幅度来讲，非线性模型效果提升幅度远高于线性模型的提升幅度。此外，二项分布统计检验结果表明，引入新闻情感的模型预测效果显著优于基准模型。

5.2 研究展望

本文研究了基于新闻情感的国际原油价格走势预测问题，以 WTI 和路透石油新闻作为研究对象，通过格兰杰因果检验确定了石油价格与新闻情感之间存在格兰杰因果关系，并通过线性与非线性预测模型确定新闻情感的引入可以提升油价走势的预测准确率，但是仍然有一些待完善的地方需要深入的研究以待解决，具体如下：

1. 社交媒体的兴起产生了大量的社交数据，因此大量的社交数据包含了市场参与者的情感和态度，社交数据可以充分反映市场变化规律，本文使用的是新闻数据并没有引入社交数据，因此下一步工作就是使用海量社交数据辅助市场预测。

2. 虽然新闻情感可以简单提取新闻中的信息，但是提取的信息是很少量的，而随着深度学习在自然语言处理领域的发展，深度学习将会从文本信息中提取更多有效的信息特征，因此本文下一步工作希望借助深度学习展开石油价格预测工作。

万方数据

万方数据

参考文献

[1] Hubbard N E, Lim D, Erickson K L. Alteration of murine mammary tumorigenesis by dietary enrichment with n-3 fatty acids in fish oil[J]. Cancer letters, 1998, 124(1): 1-7

[2] Blanchard O J, Gali J. The Macroeconomic Effects of Oil Shocks: Why are the 2000s so different from the 1970s?[R]. National Bureau of Economic Research, 2007.

[3] Das S R, Chen M Y. The western texas Intermediate (WTI) crude oil price[J]. Management Science, 2007, 53(9), 1375-1388.

[4] Lo S F, Lu W M. Does size matter? Finding the profitability and marketability benchmark of financial holding companies[J]. Asia-Pacific Journal of Operational Research, 2006, 23: 229-246

[5] Ji C, Li Y, Qiu W, Jin Y. Big data processing: Big challenges and opportunities[J]. Journal of Interconnection Networks, 2012, 13:1250009

[6] Xu J, Huang E, Chen CH. Simulation Optimization: A Review and Exploration in the New Era of Cloud Computing and Big Data[J]. Asia-Pacific Journal of Operational Research, 2015, 1550019.

[7] Knetsch T A. Forecasting the price of crude oil via convenience yield predictions[J]. Journal of Forecasting, 2007, 26: 527-549

[8] Ghosh, S. Import demand of crude oil and economic growth: evidence from India. Energy Policy, 2007, 37: 699-702

[9] Akarca A T, Andrianacos D. Detecting break in oil price series using the Box-Tiao method[J]. International Advances in Economic Research, 1997, 3(2): 217-224.

[10] Lanza A, Manera M, Giovannini M. Modeling and forecasting cointegrated relationships among heavy oil and product prices[J]. Energy Economics, 2005, 27(6), 831-848.

[11] Lu X, Kawai K I, Maekawa K. Estimating Bivariate Garch-Jump Model Based On High Frequency Data: The Case Of Revaluation Of The Chinese Yuan In July 2005[J]. Asia-Pacific Journal of Operational Research, 2010, 27(02): 287-300.

[12] Escobar M, Olivares P. Risk Management Under A Factor Stochastic Volatility Model[J]. Asia-Pacific Journal of Operational Research, 2011, 28(01): 65-80.

[13] Mirmirani S, Li H C. A comparison of VAR and neural networks with genetic algorithm in forecasting price of oil[J]. Advances in Econometrics,2004, 19: 203-223.

[14] Li J P, Tang L, Sun X L. Country risk forecasting for major oil exporting countries: A decomposition hybrid approach[J]. Computers and Industrial Engineering, 2012, 63(3): 641-651.

[15] Tang L, Yu L, Wang S. A novel hybrid ensemble learning paradigm for nuclear energy consumption forecasting[J]. Applied Energy, 2012, 93: 432-443

[16] Oladosu G. Identifying the oil price–macroeconomy relationship: An empirical mode decomposition analysis of US data[J]. Energy Policy, 2009, 37(12): 5417-5426

万方数据

[17] King K, Deng A, Metz D. An Econometric Analysis of Oil Price Movements: The Role of Political Events and Economic News, Financial Trading, and Market Fundamentals[D]. Bates White Economic Consulting. 2012.

[18] Kulkar S, Haidar I. Forecasting model for crude oil price using artificial neural networks and commodity future prices[J]. Int. J. Comp. Sci. Inf. Secur, 2009, 2(1).

[19] Khashman A, Nwulu N I. Intelligent prediction of crude oil price using Support Vector Machines[C]. Applied Machine Intelligence and Informatics (SAMI), 2011 IEEE 9th International Symposium on. IEEE, 2011: 165-169.

[20] Yu L, Zhao Y, Tang L. A compressed sensing based AI learning paradigm for crude oil price forecasting[J]. Energy Economics, 2014, 46: 236-245.

[21] Antweiler W, Frank M Z. Is all that talk just noise? The information content of internet stock message boards[J]. The Journal of Finance, 2004, 59(3): 1259-1294.

[22] Wuthrich B, Cho V, Leung S W. Daily stock market forecast from textual web data[C]. Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on. IEEE, 1998, 3: 2720-2725.

[23] Fung G P C, Yu J X, Lam W. Stock prediction: Integrating text mining approach using real-time news[C]. Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on. IEEE, 2003: 395-402.

[24] Chatrath A, Miao H, Ramchander S. Currency jumps, cojumps and the role of macro news[J]. Journal of International Money and Finance, 2014, 40: 42-62.

[25] Rachlin G, Last M, Alberg D. ADMIRAL: A data mining based financial trading system[C]. Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on. IEEE, 2007: 720-725.

[26] Schumaker R P, Zhang Y, Huang C N. Evaluating sentiment in financial news articles[J]. Decision Support Systems, 2012, 53(3): 458-464.

[27] Huang C J, Liao J J, Yang D X, et al. Realization of a news dissemination agent based on weighted association rules and text mining techniques[J]. Expert Systems with Applications, 2010, 37(9): 6409-6413.

[28] Peramunetilleke D, Wong R K. Currency exchange rate forecasting from news headlines[J]. Australian Computer Science Communications, 2002, 24(2): 131-139.

[29] Bollen J, Mao H. Twitter mood as a stock market predictor[J]. Computer, 2011, 44(10): 91-94.

[30] Mahajan A, Dey L, Haque S M. Mining financial news for major events and their impacts on the market[C]. Web Intelligence and Intelligent Agent Technology, 2008. IEEE/WIC/ACM International Conference on. IEEE, 2008, 1: 423-426

[31] Schumaker R P, Chen H. Textual analysis of stock market prediction using breaking financial news: The AZFin text system[J].ACM Transactions on Information Systems (TOIS),2009, 27(2): 12.

[32] Zhai Y, Hsu A, Halgamuge S K. Combining news and technical indicators in daily stock

万方数据

price trends prediction[M]. Advances in Neural Networks: Springer Berlin Heidelberg, 2007: 1087-1096.

[33] Vu T T, Chang S, Ha Q T. An experiment in integrating sentiment features for tech stock prediction in twitter[J]. 2012

[34] Hagenau M, Liebmann M, Neumann D. Automated news reading: Stock price prediction based on financial news using context-capturing features[J]. Decision Support Systems, 2013, 55(3): 685-697.

[35] Jin F, Self N, Saraf P, et al. Forex-foreteller: Currency trend modeling using news articles[C]. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013: 1470-1473.

[36] Butler M, Kešelj V. Financial forecasting using character n-gram analysis and readability scores of annual reports[M]. Advances in artificial intelligence. Springer Berlin Heidelberg, 2009: 39-51

[37] Mittermayer M A, Knolmayer G. Text mining systems for market response to news: A survey[M]. Institut für Wirtschaftsinformatik der Universität Bern, 2006

[38] Mittermayer M A, Knolmayer G F. Newscats: A news categorization and trading system[C]. Data Mining, 2006. ICDM'06. Sixth International Conference on. Ieee, 2006: 1002-1007.

[39] Devitt A, Ahmad K. Sentiment Polarity Identification in Financial News: A Cohesion-based Approach[J].ACL 2007, 984.

[40] Lechthaler F, Leinert L. Moody Oil-What is Driving the Crude Oil Price?[J]. 2012. CER-ETH-Center of Economic Research (CER-ETH) at ETH Zurich.

[41] Pröllochs N, Feuerriegel S, Neumann D. Generating Domain-Specific Dictionaries Using Bayesian Learning[J]. Available at SSRN 2522884, 2014.

[42] Alfano S J, Feuerriegel S, Neumann D. Is News Sentiment More than Just Noise?[J]. Available at SSRN, 2014.

[43] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and trends in information retrieval, 2008, 2(1-2): 1-135.

[44] Minev M, Schommer C, Grammatikos T. News and stock markets: A survey on abnormal returns and prediction models[R]. Technical Report, UL, 2012.

[45] Nassirtoussi A K, Aghabozorgi S, Wah T Y. Text mining for market prediction: A systematic review[J]. Expert Systems with Applications, 2014, 41(16): 7653-7670.

[46] Demers E, Vega C. Soft information in earnings announcements: News or noise?[M]. Federal Reserve Board, 2008.

[47] Henry E. Are Investors Influenced By How Earnings Press Releases Are Written?[J]. Journal ofBusiness Communication. 2008, 45 (4): 363–407.

[48] Jegadeesh N, Wu D. Word power: A new approach for content analysis[J]. Journal of Financial Economics, 2013, 110(3): 712-729.

[49] Loughran T, McDonald B. When is a liability not a liability? Textual analysis,

万方数据

dictionaries, and 10 - Ks[J]. The Journal of Finance, 2011, 66(1): 35-65

[50] Tetlock P C, SAAR - TSECHANSKY M, Macskassy S. More than words: Quantifying language to measure firms' fundamentals[J]. The Journal of Finance, 2008, 63(3): 1437-1467.

[51] Antweiler W, Frank M Z. Is all that talk just noise? The information content of internet stock message boards[J]. The Journal of Finance, 2004, 59(3): 1259-1294.

[52] Li J P, Tang L, Sun X L, et al. Country risk forecasting for major oil exporting countries: a decomposition hybrid approach [J]. Computers & Industrial Engineering, 2012, 63(3): 641-651.

[53] Groth S S, Muntermann J. An intraday market risk management approach based on textual analysis[J]. Decision Support Systems, 2011, 50(4): 680-691.

[54] 侯璐. 基于 ARIMA 模型的石油价格短期分析预测[D]. 广州：暨南大学, 2009.

[55] 肖龙阶, 仲伟俊. 基于 ARIMA 模型的我国石油价格预测分析[N]. 南京航空航天大学学报.2009. 11.4.41-46.

[56] 丁静之, 闵骐, 林怡. ARIMA 模型在石油价格预测中的应用[J]. 物流技术, 2008, 27(10): 156-159.

[57] Yu L, Zhao Y, Tang L. A compressed sensing based AI learning paradigm for crude oil price forecasting [J]. Energy Economics, 2014, 46: 236-245.

[58] Xie W, Yu L, Xu S, et al. A new method for crude oil price forecasting based on support vector machines [M]//Computational Science–ICCS 2006. Springer Berlin Heidelberg, 2006: 444-451.

[59] 贾振华, 陈英杰. 神经网络在石油价格预测中的仿真研究[J]. 计算机仿真, 2011, 28(11), 354-357.

[60] 王欣冉, 邢永丽, 巨程晖. 小波包与贝叶斯 LS-SVM 在石油价格预测中的应用[J]. 统计与决策, 2011, 6, 162-164.

[61] 韩春, 田大钢. 对股票市场信息的文本挖掘[J]. 中国高新技术企业, 2008, 23: 6-8.

[62] 赵丽丽.互联网财经新闻对股市影响的定量分析[D]. 成都：西南财经大学, 2012.

[63] 陈茜,连婉琳. 基于文本挖掘技术的互联网股票新闻的情感分类[J]. 中国市场, 2015, 24: 234-235.

[64] 张世军. 基于网络舆情的 SVM 股票价格预测研究[D]. 南京：南京信息工程大学,2014.

[65] 黄润鹏, 左文明, 毕凌燕. 基于微博情绪信息的股票市场预测[N]. 管理工程学报, 2015. 1.47-52.

[66] 彭敏, 汪清, 黄济民. 基于情感分析技术的股票研究报告分类[N]. 武汉大学学报: 理学版, 2015.2.124-130.

[67] Alvarez-Ramirez J, Soriano A, Cisneros M, et al. Symmetry/anti-symmetry phase transitions in crude oil markets [J]. Physica A: Statistical Mechanics and its Applications, 2003, 322: 583-596.

[68] Shin H, Hou T, Park K, et al. Prediction of movement direction in crude oil prices based

万方数据

on semi-supervised learning[J]. Decision Support Systems, 2013, 55(1): 348-358.

[69] Liebmann M, Hagenau M, Neumann D. Information Processing in Electronic Markets: Measuring Subjective Interpretation Using Sentiment Analysis[J]. 2012.

[70] Granger C W J. Investigating causal relations by econometric models and cross-spectral methods [J]. Econometrica: Journal of the Econometric Society, 1969: 424-438.

[71] Cox D R. Planning of experiments [J]. 1958.

[72] Kumar P R, Ravi V. Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review[J]. European journal of operational research, 2007, 180(1): 1-28.

[73] Quinlan J R. Induction of decision trees[J]. Machine learning, 1986, 1(1): 81-106.

[74] Rumelhart D E, Smolensky P, McClelland J L, et al. Sequential thought processes in PDP models[J]. V, 1986, 2: 3-57.

[75] Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297

[76] Nochai R, Nochai T. ARIMA model for forecasting oil palm price[C]//IMT-GT regional conference on Mathematics. Statistics and Applications. Universiti Sains Malaysia, Penang June-13-15. 2006.

[77] Mohammadi H, Su L. International evidence on crude oil price dynamics: Applications of ARIMA-GARCH models[J]. Energy Economics, 2010, 32(5): 1001-1008.

万方数据

万方数据

致谢

首先感谢我的导师李健教授对我的培养和教育，应该说是李老师当初对我的肯定和认可才使得我开始了学术研究之路，李老师的对学术的热爱和执着深深影响了我，我开始喜欢上学术研究，认识到学术研究的本质其实是和做人是一样的，都要踏踏实实、勤勤恳恳，做学问是寂寞的，但也是有价值的，感谢老师让我找到人生中有意义有价值的事情，我想我会一直坚持下去。

由于研究生期间我是跟随汤铃教授和余乐安教授做数据挖掘和文本挖掘领域的研究，所以真诚地感谢余老师和汤老师的批评和指导，可以说现在我身上的所有的本领都是从两位老师身上学来，无论是刻苦研究的精神还是科学严谨的研究态度深深地影响着我学习和工作，当然我也会永远记得汤老师的座右铭“态度决定人生、细节决定高度”，我也会铭记余老师的批评教育，认认真真做事，勤勤恳恳做人。

再次，还要感谢所有经济管理学院的老师们，感谢老师您无私的教导和培养，我会谨记各位老师的教诲，在学习的路上继续前行。

研究生期间最大的财富莫过于和同门师兄、师姐、师弟、师妹的快乐时光，也希望我们未来的日子也常常聚在一起，像兄弟姐妹一样。我也会铭记大家对我的帮助，感谢生命中有你们这么一群“小伙伴”。

徐振敬

2016 年 5 月

万方数据

万方数据

研究成果及发表学术论文

已撰写的论文

Online-Purchasing Behavior Forecasting with a Hybrid Firefly Algorithm based SVM Model Considering Shopping Cart Use. *European Journal of Operational Research.*（在投）

Forecasting oil price trends with sentiment of online news articles. *Asia-Pacific Journal of Operational Research.*（在投）

参与项目

中国石油天然气股份有限公司规划总院项目：国际油价短期预测模型开发(2014.03-2015.03)

万方数据

万方数据

作者和导师简介

作者简介：徐振敬，男，汉族，1990 年 7 月出生于山东省临清市。2009 年考入青岛大学国际商学院信息管理与信息系统专业，2013 年获得管理学学士学位。同年考入北京化工大学经济管理学院管理科学与工程专业，攻读工学硕士学位，师从李健教授和汤铃副教授，研究文本挖掘与情感分析的先关工作，论文工作进展顺利。

李健，男，1976 年 3 月 5 日，汉，山东泰安人，中国科学院数学与系统科学研究院管理学博士，加拿大温莎大学访问学者，现为北京化工大学经济管理学院教授，博士生导师，2012 年入选教育部“新世纪优秀人才支持计划”。目前担任中国系统工程学会监事会监事、中国运筹学学会决策科学分会常务理事、北京运筹学会理事、国际期刊 International Journal of Inventory Research 编委。

研究方向：物流与供应链管理。到目前为止，发表论文 40 余篇（SCI 检索 10 余篇，EI 检索 10 余篇），由 Springer 出版英文专著 2 部。联系方式：lijian@mail.buct.edu.cn

万方数据

北京化工大学

硕士研究生学位论文答辩委员会决议书

研究生姓名：徐振敬　　专业：管理科学与工程

论文题目：基于情感分析的国际原油价格走势预测研究

指导教师姓名：李健　　职称：教授

论文答辩日期：2016 年 5 月 22 日　　地点：化纤楼 207

论文答辩委员会成员

姓名	职称	工作单位	本人签名
方勇	教授	北京化工大学经管学院	
刘斌	副教授	北京化工大学经管学院	
任继勤	副教授	北京化工大学经管学院	
吴卫红	副教授	北京化工大学经管学院	
张爱美	副教授	北京化工大学经管学院	

注：此表用于存档，除本人签名务必用钢笔填写外，其余处必须用计算机打印。

万方数据

答辩委员会对论文的评语（选题意义、文献综述、论文所取得的成果及水平、学风和论文写作水平、论文的不足之处）：

徐振敬同学的硕士论文《基于情感分析的国际原油价格走势预测研究》选题合理，具有较强的理论及现实意义。

该论文作者在查阅大量相关文献的基础上，对石油价格的复杂性和非平稳性进行了归纳分析，对比研究了国内外石油价格预测方法，文献综述较为清晰全面。本文提出了基于情感分析的国际原油价格走势预测模型，就石油相关新闻的分析，采用领域关键词词典的方法，得到新闻的情感序列，再通过格兰杰因果检验法，得到情感序列和石油价格序列的相关性和滞后期，最后通过机器学习的方法预测石油价格走势，并选取美国西德克萨斯轻质原油期货价格和路透社原油新闻中新闻情感作为研究对象分析石油价格预测能力，发现存在较高的正相关关系，新闻情感变化早于石油价格变化，新闻情感值的正负和变化同时影响石油价格的变化。论文不足之处在于：第一，论文中某些逻辑关联词语表述不够严谨；第二，论文阐述描写不够详实；第三，论文排版中存在少量格式错误。

该论文立意较好，主题突出，结构合理，文献较为丰富，思路清晰，语言较为严谨，达到了硕士研究生水平。此外，徐振敬同学答辩表达比较清晰，准确流利地回答了评委提出的问题。

同意授予徐振敬同学硕士学位。报院学位委员会讨论。

对学位论文水平的总体评价	优秀	良好	一般	较差
		√		

答辩委员会表决结果：

同意授予硕士学位 5 票，不同意授予硕士学位 0 票，弃权 0 票。根据投票结果，答辩委员会做出建议授予该同学硕士学位的决议。

答辩委员会主席签字：

2016 年 5 月 22 日

万方数据