信 息 科 学 与 技 术 学 院

毕 业 设 计 开 题 报 告

面向网络股评观点的垂直搜索引擎的设计与实现

学 号: 2014112200

姓 名: 张浩天

班 级: 4班

专 业: 软件工程

指导老师: 陶宏才

职 称: 教 授

2018年1月17日

开题报告的内容应包括

1. 本课题的目的、意义。
2. 本课题国内外研究现状（国内外文献综述，给出参考文献）。
3. 本课题设计任务与要求。
4. 拟采取的技术路线与试验方案。
5. 预期成果（包括预期能够完成的设计或者理论研究成果，拟提交的软件、硬件、仿真程序等）。
6. 设计进度安排。

注：（1）开题报告工作由毕业设计指导小组组织实施，目的是帮助学生执行选题论证。

（2）毕业设计指导小组的论证意见以“通过”、“不通过”结论。通过者按计划开展论文工作。不通过者，整改后重新提交开题报告，如果仍然不通过则取消毕业设计资格。

|  |
| --- |
| 开 题 报 告 主 要 内 容   1. 本课题的目的、意义   股市对国家经济发展起到非常重要的作用，我国自上个世纪九十年代以来先后在上海市和深圳市建立了证券交易所，经过二十多年来不断的发展，站在全球第二大经济体的肩膀上，中国股票A股市场已经成为是全球第二大资本市场，也是全球第一大新兴资本市场，还是全球成长速度最快的资本市场之一。随着股票市场的不断发展和股票市场对投资者的影响不断加深，越来越多的学术与金融届从业者也在探索尝试股市行情趋势的分析与研究，从而能够实现对股票的走势进行预测。  网民对股市或股票的网络评论观点在很大程度上反映了股市行情，也影响着股市涨跌。因此，快速高效地挖掘出网民对股市或股票的态度与观点，对股市预测具有很大的指导意义。目前很多研究方法都已经成功证明，股民的股票操作决策很大程度上都到网络股民观点和感情因素的影响，但这些研究成果都具有一定的局限性。综上，设计一个面向网络股评观点的垂直搜索引擎是一个值得研究的方向，通过融合全文搜索和股评观点挖掘，实现在线股评观点的情感极性分析，并且能够检索网络股评内容，在不同指标下满足实际应用的要求。   1. 本课题国内外研究现状   近年来，越来越多的研究者与金融业从业者在研究各种股票的预测方法。第40次《中国互联网络发展状况统计报告》指出，中国网民规模达7.51亿，互联网普及率达到54.3%[1]。针对网络股评是否会影响股票市场与股市行情行情的问题，张对通过收集网络股评并提取情感指数，认为从网络股评中提取投资者观点可行并具有显著优点[2]。宋敏晶通过文本分析投资者的股评行为，利用主成分回归股票价格预测模型来实证其影响[3]。因此，网民在网络上对整个股票市场或者是对某个特定的股票的评论观点在很大程度上能够反映着市场或者是股票的行情。  JOHAN B等基于网络数据，提出了利用神经智能网络对股票预测的方法[4]。赵丽丽通过使用计量经济学的分析与计算机科学的跨学科方法，利用本文挖掘技术，将财经新闻量化成对股市影响的一个因子，经过多元回归分析确定了互联网财经新闻会对股票价格产生的具体影响[5]，但其只研究了新闻本体信息，而没有分析对应的网民评论信息。对此，徐琳针对网络上重大突发性舆情事件有关的微博信息、评论等数据进行分析，对比涉事股票的行情表现与其他特征证实了利空与利好类型舆情事件均对股票市场具有影响[6]。文献[7]通过使用网络搜索引擎获得具有较高语义相似度的短文本，用于金融领域从而可以预测股票的相关变化。文献[8]选用网络证券分析师的股票预测信息，通过探究个股股评数量、收益率等与股票价格波动率的变动，发现了互联网股评信息对证券市场风险的影响。  面对浩瀚互联网信息，寻找到需要的信息成为了一个重要问题，传统搜索引擎的的搜集速度远远赶不上信息的增长速度，查询到的结果中也存在大量重复和垃圾信息。而垂直搜索引擎可以面向某一另据收录专业化的信息[9]。垂直搜索引擎也在不断发展，近年来也运用到多个领域。文献[10]和文献[11]对基于Lucene的垂直搜索引擎进行了基本的探索，使用爬虫工具抓取有关页面，对网页进行分析进行信息提取后建立倒排序索引模块，最终实现信息查询功能。但是没有深入研究有关的技术，在索引和结果的查询及呈现仍然有改进的余地。  主题爬虫技术是垂直搜索引擎的重要组成部分，关系着最终搜索的结果的质量。文献[12]提出利用贝叶斯分类器对网页进行分类的方法，这种方法更适用在网络资源数量较大的情况。文献[13]提出了基于内容的空间向量模型和遗传算法相结合的搜索策略，可以保证在初始阶段的爬虫抓取效率，仍有较大改善空间。文献[14]基于内容评价策略，提出鱼搜索算法，文献[15]在此基础上进行了改进出了鲨鱼搜索算法，从而更精确的划分爬取的优先级并计算之间的相关性。文献[16]根据链接的价值作为网页优先性的重要依据，提出了PageRank（网页排名）算法并被许多搜索引擎采用，文献[17]提出了基于链接结构的PageRank算法，改善了该算法出现的主题偏移问题。  在情感分析方面，文献[18]通过开发情感分析工具，对网络文章与评论信息与股票表现进行比对，使投资者在市场中获取更好的效益。林炳灿利用网络舆论信息中的情感信息，针对现有研究基础薄弱与信息利用不足的问题，引入不同因素模型与投资者情绪模型，采用主观方法利用软件测度投资者情绪，对股票价格进行研究[19]。文献[20]和文献[21]首先提取股评文本信息利用SVM（支持向量机）机器学习实现了文本情感分类，同时利用一种基于网络舆情和股票技术指标的支持向量机回归模型从而预测股票收盘价格，相对于传统的SVM回归模型显示出了较高的精确度。莫倩基于股评文章的特征，利用基于模式的倾向性股评分析方法，识别提取倾向性观点，实现了针对股评观点的挖掘系统[22]。  网络股评信息众多，鱼龙混杂，在如此巨大的数据中，同时存在着无用的垃圾信息与虚假意见，它们的制造和传播扰乱了网络秩序[23]，也困扰着我国股票市场的健康发展。针对股评文章与一般网络评论的不同，文献[24]将基于篇章结构的股评观点倾向性分析方法和基于模式的股评倾向性分析方法相结合，找出并去除了偏离正常股评的不真实信息，综合成一种混合的股评观点倾向性分析方法，提高了分析的准确率。胡航丽等提出了利用篇章结构改进股评观点分类，提取股评的标题和预测性语句，在SVM算法的基础上对股评进行了更好的分类预测，也根据其评论特征发现了潜藏的网络水军[25]。  以上针对股票市场趋势的预测的研究成果都存在一定的局限，而文献[26]运用面向股市的垂直搜索引擎，并通过收集博客评论等信息验证了其与股票市场变化的相关性，同时取得了较好的效果。因此，融合股评观点挖掘的面向股评的垂直搜索引擎是一个值得研究的方向。  **参考文献**   1. 中国互联网信息中心. 第40次《中国互联网络发展状况统计报告》[DB/OL]. http: //www.cnnic.net.cn, 2018年1月7日 2. 张对. 网络股评影响股市走势吗——基于股票情感分析的视角[J]. 现代经济信息, 2015(01): 355-357. 3. 宋敏晶. 基于情感分析的股票预测模型研究[D]. 硕士论文. 哈尔滨工业大学, 2013. 4. JOHAN B, Mao Huina, Zeng Xiaojun. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2011, 2(1) : 1-8 5. 赵丽丽. 互联网财经新闻对股市影响的定量分析[D]. 硕士论文. 西南财经大学, 2012. 6. 徐琳. 网络舆情对股价波动影响的实证研究[D]. 硕士论文. 西南财经大学, 2013. 7. HUANG B, HEILMAN T D. A web-based kernel function for measuring the similarity of short text snippets[C]. Proceedings of the 16th International Conference on World Wide Web. ACM, 2007: 377-386. 8. 邵凯, 童亮, 段江娇. 互联网股评信息对股票价格波动的影响分析[J]. 改革与开放, 2017(13): 13-15. 9. 王文钧, 李巍. 垂直搜索引擎的现状与发展探究[J]. 情报科学, 2010, 28(03): 477-480. 10. 方志民, 戴洋洋, 董淑珍, 李渤, 温芳馨, 宋新航. 新闻类垂直搜索引擎系统研究与设计[J]. 黑龙江工程学院学报, 2016, 30(06): 35-37. 11. 李全, 林松, 田俊, 刘兴红. 面向MOOC的垂直搜索引擎[J]. 计算机与现代化, 2017(04): 32-37. 12. 邹永斌, 陈兴蜀, 王文贤. 基于贝叶斯分类器的主题爬虫研究[J]. 计算机应用研究, 2009, 26(09): 3418-3420+3439. 13. 张小琴, 王晓辉. 主题信息搜索系统中的搜索策略研究[J]. 软件导刊, 2014, 13(01): 89-92. 14. De Bra P, Houben G J, Korn atzky Y, et al. Information Retrieval in Distributed Hypertexts[C]. RIAO 1994, 4th International Conference, NY, USA, 1994, 481-493 15. L. Luo, Q. Chen, Q. Wu. Research on topical crawler of shark-search algorithm and hits algorithm. Computer Technology and Development, 2010, 11: 02 16. D. R Gleich. PageRank beyond the Web. SIAM Review, 2015, 57(3): 321-363. 17. Taher H. Haveliwala. Topic Sensitive PageRank: a context-sensitive ranking algorithm for Web search[J]. IEEE Transactions on Knowledge & Data Engineering, 2003, 15(4): 784-796 18. Wang Gang, Wang Tianyi, Wang Bolun, et al. Crowds on wall street: Extracting value from collaborative investing platforms[C]. In Progress in CSCW, 2015. 19. 林炳灿. 基于投资者情绪的网络舆论对股票价格影响的统计研究[D]. 硕士论文. 西南财经大学, 2013. 20. 张世军. 基于网络舆情的SVM股票价格预测研究[D]. 硕士论文. 南京信息工程大学, 2014. 21. 张世军, 程国胜, 蔡吉花, 杨建伟. 基于网络舆情支持向量机的股票价格预测研究[J]. 数学的实践与认识, 2013, 43(24): 33-40. 22. 莫倩, 姜越, 胡航丽. 基于web的股评观点挖掘系统[J]. 计算机系统应用, 2012, 21(12): 38-42+51. 23. 莫倩, 杨珂. 网络水军识别研究[J]. 软件学报, 2014, 25(07): 1505-1526. 24. 莫倩, 张渝杰, 胡航丽, 张华平. 一种混合的股评观点倾向性分析方法[J]. 计算机工程与应用, 2011, 47(19): 222-225. 25. 胡航丽, 莫倩. 利用篇章结构改进股评观点分类的研究[J]. 小型微型计算机系统, 2009, 30(05): 899-902. 26. CHOUDHURY M M, SUNDARAM H, JOHN A. Can blog communication dynamics be correlated with stock market activity[C]. In Proc. of HyperText, 2008: 55-60. 27. 本课题设计任务与要求   面向网络股评观点的垂直搜索引擎应该具备的主要功能包括：  （1）定点收割爬虫模块（面向特定网站的主题数据采集和属性统计模块）：将目录搜索技术和主题爬虫技术相结合，由用户来指定一些金融网站的信息，设定对应的定时参数从而爬取这些网站中的数据。将数据采集后会进行清洗解析和结构化的抽取步骤，获取用户需要的数据。  （2）情感极性分类模块：通过考虑语句与短语等对情感词极性的影响，分析情感极性与强度。采用模糊集合理论对其进行分类，在无监督情况下自我学习，构成情感分类器。由情感知识库模块、情感强度的多粒度计算模块和模糊分类器模块构成。  （3）文件生成与存储模块：融合情感极性分类和关键词索引功能，生成倒排序文件并完成数据的保存管理和标注。  （4）文件搜索模块：提供用户对股评的全文检索以及情感极性和相关属性条件的组合统计查询，包括文件的检索、查询与呈现。  如图1所示为软件概要系统功能设计图。    图1 系统功能设计图   1. 拟采取的技术路线与试验方案   面向网络股评观点的垂直搜索引擎是基于apache的Lucene改进来实现的。Lucene是一套用java实现的用于全文检索和搜索的开放源代码程序库，在此基础上来增加核心模块，并对已有的模块进行功能扩展和性能优化，具体试验方案如下：  （1）理论分析：根据目前现有的由网络股评预测股票信息等研究成果，了解目前技术发展，分析相关的分析流程、方法，优点与不足。与此同时，充分理解垂直搜索引擎在近年来的发展，实现方法与具体内容。在掌握原理以及历年的基础上，研究面向股评观点的垂直搜索引擎，分析其具体实现方法与相关依赖平台。  （2）设计制作：学习爬虫技术，在现有爬虫技术的基础上，设计一种可以面向指定的网站进行定时收割的爬虫基础，将目录搜索技术和主题爬虫技术结合。支持用户对网站和内容进行自助定制，并由此设置爬虫的相关参数。构建有关的URL（统一资源定位符）列表进行网页抓取，将收集到的数据根据目录列表和主题描述进行文本过滤，只获取用户需要的有关数据。  设计一个多粒度的支持模糊计算的无监督股评极性情感分类方法，为了提高分类的准确度和精度，需要针对情感词语在不同层次上对上下文的情感信息进行分析，并在结合模糊集合理论的基础上构建一个分类模型。设计完成无监督情感极性分类器的情感知识库、股评情感强度的多粒度计算组件和模糊分类器三部分。  掌握apache Lucene的设计架构，按照软件工程开发流程进行系统的设计开发工作。首先能够正常使用Lucene，在此基础上改进代码，增加上述两个面向特定网站的主题数据采集和属性统计模块、情感极性分类模块。对Lucene索引创建，索引和搜索部分进行改进，对排序生成和文件检索模块、查询及呈现模块进行功能扩展和性能优化，根据实际情况对模块进行改进。  （3）实验运行：系统初步完成后，先在本地进行测试实现系统的实际使用情况，观察各个模块的运行情况，及时进行改善。测试系统的安全性以及可靠性，根据实现的实际情况，分析发现存在的问题。根据存在的问题进行改进设计，重新进行调试运行，直到满足设计的任务与要求为止。整理相关开发文档，完成毕业设计论文。   1. 预期成果   毕业设计论文；  英文文献翻译；  面向网络股评观点的垂直搜索引擎设计源代码。   1. 设计进度安排   （1）第一阶段：毕业设计准备阶段  第七学期  第17-18周：查看毕业设计任务书，理解该毕业设计的主要内容及任务；  查阅相关文献资料，了解相关知识，准备开题报告的撰写；  完成所需开发工具与平台的准备工作。  第 19 周：整理文献资料，撰写开题报告；  了解所采用的技术，平台软件的学习与使用。  第 20 周：准备开题答辩，提交开题报告，进行开题答辩。  （2）第二阶段：毕业设计的设计阶段  第八学期  第 1 周：针对项目进行需求分析，准确敲定项目的功能需求；  撰写需求分析书。  第 2 周：根据需求进行项目的概要设计和详细设计；  撰写项目的概要设计说明书和详细设计说明书。  （3）第三阶段：项目的编码阶段  第 3 – 4 周：进行数据库的设计，编码绘制相应的表、视图等；  进行软件界面设计，编码绘制界面；  进行数据库的连接；  学会熟练利用Lucene进行开发。  第 5 – 6 周：实现情感极性分类、面向特定网站的主题数据采集和属性统计等模块；  针对倒排序生成和文件检索、查询及呈现等模块的功能扩展与性能优化；  基本完成整体功能。  （4）第四阶段：项目的测试阶段  第 7 周：整合系统，进行安全性、可靠性设计与测试，系统改进；  准备中期检查。  （5）第五阶段：毕业设计论文撰写阶段  第 8 周：整理前期毕业设计准备资料和过程记录笔记，准备撰写论文资料；  拟定毕业论文提纲，编写章节与目录编排的大致内容。  第 9 -10 周：资料整理，论文撰写与修改。  第 11 周：提交论文完整版初稿。  第 12 周：论文修改与定稿，论文装订，成果演示录屏，准备答辩演示文稿。  第 13 周：评阅及答辩，整改论文，正稿打印与存档。 |
|  |
| 指导教师意见    指导教师签字： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  年 月 日 |
| 毕业设计指导小组意见  （签字）        年 月 日 |

注 1、以上各项空格不够可以续页。

2、此表作为附件装入毕业设计（论文）资料袋存档。