# 文本复制检测报告单(全文标明引文)

ADBD2017R_20170505201619424730312336　　检测时间：2017-05-05 20:16:19

检测文献:　基于网络爬虫的股票信息预警系统的设计与实现

作者:　刘明钧

检测范围:

中国学术期刊网络出版总库
中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库
中国重要会议论文全文数据库
中国重要报纸全文数据库
中国专利全文数据库
互联网资源(包含贴吧等论坛资源)
英文数据库(涵盖期刊、博硕、会议的英文数据以及德国Springer、英国Taylor&Francis 期刊数据库等)
港澳台学术文献库
优先出版文献库
互联网文档资源
图书资源
CNKI大成编客-原创作品库
大学生论文联合比对库
个人比对库

时间范围:　1900-01-01至2017-05-05

指导教师　闫昱

## 检测结果

总文字复制比：4.6%　　跨语言检测结果：0%

去除引用文献复制比：4.6%　　去除本人已发表文献复制比：4.6%

单篇最大文字复制比：0.9%（珠宝管理销售系统软件建模与分析.doc 20页-高清全文免费预览-max文档）

| | | |
|---|---|---|
| 重复字数：[988] | 总字数：[21594] | 单篇最大重复字数：[190] |
| 总段落数：[2] | 前部重合字数：[128] | 疑似段落最大重合字数：[605] |
| 疑似段落数：[2] | 后部重合字数：[860] | 疑似段落最小重合字数：[383] |

指　标：☑疑似剽窃观点　☑疑似剽窃文字表述　☐疑似自我剽窃　☐疑似整体剽窃　☐过度引用

表格：2　　脚注与尾注：0

3.9%（383）　基于网络爬虫的股票信息预警系统的设计与实现_第1部分（总9718字）

5.1%（605）　基于网络爬虫的股票信息预警系统的设计与实现_第2部分（总11876字）

（注释：无问题部分　文字复制比部分　引用部分）

## 疑似剽窃观点　(1)

基于网络爬虫的股票信息预警系统的设计与实现_第1部分

1.　因此，从经济角度来看，本系统的开发具备经济可行性。

## 1. 基于网络爬虫的股票信息预警系统的设计与实现_第1部分　　总字数：9718

相似文献列表　文字复制比：3.9%(383)　疑似剽窃观点：(0)

| | | |
|---|---|---|
| 1 | 软件1101-1111610713-郑步健<br>郑步健 - 《大学生论文联合比对库》- 2015-05-18 | 0.9%（86）<br>是否引证：否 |
| 2 | 11071226-陈露-面向Web的天然气气质数据管理系统<br>陈露 - 《大学生论文联合比对库》- 2015-06-08 | 0.7%（69）<br>是否引证：否 |

| | | |
|---|---|---|
| 3 | 网络数据资源自动获取技术研究与应用<br>牛敏;米石云;张倩; -《信息技术》- 2013-12-25 | 0.7%（64）<br>是否引证：否 |
| 4 | 信息工程学院资源保障信息系统设计与实现<br>李赛 -《大学生论文联合比对库》- 2016-05-04 | 0.4%（43）<br>是否引证：否 |
| 5 | 家庭账务管理系统<br>杨洋 -《大学生论文联合比对库》- 2016-05-03 | 0.4%（40）<br>是否引证：否 |
| 6 | 员工考勤系统设计与实现<br>万锁 -《大学生论文联合比对库》- 2015-05-30 | 0.4%（40）<br>是否引证：否 |
| 7 | 20113345_万锁_员工考勤系统设计与实现<br>万锁 -《大学生论文联合比对库》- 2015-06-18 | 0.4%（40）<br>是否引证：否 |
| 8 | 1115240509-马占华-学生信息管理系统的设计与实现<br>马占华 -《大学生论文联合比对库》- 2015-04-30 | 0.4%（39）<br>是否引证：否 |
| 9 | 基于数据仓库的云南移动经营分析系统的应用与研究<br>汪希(导师：谢戈) -《云南大学硕士论文》- 2010-04-01 | 0.3%（32）<br>是否引证：否 |
| 10 | 金融<br>-《大学生论文联合比对库》- 2014-08-07 | 0.3%（31）<br>是否引证：否 |
| 11 | 智游图摄欣赏吧<br>王艺茹 -《大学生论文联合比对库》- 2016-05-26 | 0.3%（31）<br>是否引证：否 |
| 12 | 20122430232_王艺茹_信息工程学院_智游图摄欣赏吧<br>王艺茹 -《大学生论文联合比对库》- 2016-06-01 | 0.3%（31）<br>是否引证：否 |
| 13 | 信息科学学院_09070846_张欢<br>信息科学学院 -《大学生论文联合比对库》- 2013-05-28 | 0.3%（29）<br>是否引证：否 |

原文内容

本科毕业设计（ 论文 ）

题目： 基于网络爬虫的股票信息预警系统的设计与实现

学号： 20134830347

姓名： 刘明钧

班级： 13软工A2

专业： 软件工程

学部（院）： 工学部

入学时间 ： 2013级

指导教师 ： 闫昱

日期： 2017年05月03日

毕业设计（论文）独创性声明

本人所呈交的毕业论文是在指导教师指导下进行的工作及取得的成果。除文中已经注明的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确说明并表示谢意。

作者签名：

日期：

基于网络爬虫的股票信息预警系统的设计与实现

摘要

本文结合网络爬虫技术实现对于股票交易信息、股票公告信息、股票财务信息的采集、解析、格式化、挖掘、维护与展示。再通过用户预设的条件对抓取的信息进行推送、预警。本文通过需求分析确定了系统应具有的基本功能包括股票数据获取、页面解析、解析内容格式化、数据整理、信息维护、信息浏览、设置预警、发送预警。采用面向对象的方法进行了总体设计、详细设计并最终实现了股票信息预警系统的主要功能。

本文设计的股票信息预警系统共分为股票信息网页采集模块、网页解析模块、数据整理模块、数据浏览模块、预警模块共五个模块。股票信息采集模块采用爬虫技术实现，主要解决了如何准确快速获取增量的股票数据的问题。网页解析模块通过使用原生的XPATH 模块进行，获取需要的信息。数据处理模块采用Newtonsoft.Json库对Json 字符串对象化并存入关系型数据库。数据浏览模块是对数据库中数据的可视化展示。预警模块实现用户自我定制需要的信息条件，通过短信及邮件的方式进行推送。

目前，系统处于运营维护阶段，可以稳定、高效的进行股票数据及相关信息的采集、解析、预警。

关键词:网络爬虫；股票预警；WEB挖掘；

Design and Implementation of Stock information early warning system based on web crawler

ABSTRACT

This paper combines the web crawler technology of realizing the acquisition, analysis, formatting, excavation, maintenance and display of stock transaction information, stock announcement information and stock financial information. And then push and early warning the information crawled through the user's default conditions. This paper analyzes the basic functions that the system should have the demand analysis has defined, including stock data acquisition, page analysis, parsing content formatting, data collation, information maintenance, information browsing, setting early warning and sending early warning. The object-oriented method used to design the whole design, and the main function of the stock information early warning system finally designed and realized. The stock information early warning system designed in this paper is divided into five modules: stock information webpage acquisition module, web page analysis module, data collation module, data browsing module and early warning module. The stock information acquisition module is implemented by reptile technology, which solves the problem of how to get the incremental stock data accurately and quickly. The web analytics module makes use of the native XPATH module to get the information you need. The data processing module uses the Newtonsoft.Json library to object to the Json string and store it in a relational database. The data browsing module is a visual display of the data in the database. Early warning module achieves the user needs to customize the information conditions, and pushes it through the SMS and e-mail.At present, the system is in the stage of operation and maintenance, and stock data and related information can be collected, analyzed and warned stably and efficiently.

Key words: web crawler; stock early warning; Web mining

目录

## 1 绪论

### 1.1 研究的背景

随着我国改革开放的脚步，股票日益成为人们生活中不可或缺的投资理财工具之一。股票作为重要经济活动之一，对于国内市场经济的繁荣与国民经济的发展都起到了至关重要的作用。

伴随着互联网等相关技术的发展股票交易方式由早期的手工操作、电话委托、柜台交易直至今天的在线股票交易平台。同时用户所能获取的股票相关信息无论是在数量上还是纬度上都呈爆炸增长的趋势。用户所能获得的信息从原来的匮乏时代一下子来到了信息过载时代。因此用户需要一种手段可以对这些信息进行过滤，只获取自己关注的信息。

本文的主要目标就是通过使用爬虫技术对于股票这个垂直领域中的数据、信息进行采集、针对采集的网页进行解析、挖掘、存储。在通过用户的预设条件进行实时的信息预警。

### 1.2 研究现状

网络爬虫亦称信息采集系统是将网页中的非结构化信息进行抓取、清洗最终存入到关系型数据库中的软件。

针对股票数据具有实时更新的特点，本文采用的网络爬虫为增量采集系统。其大致的工作原理如下：

* 对所有目标网页进行抓取

* 在之后的数据抓取过程中比较原网页与新抓取网页，对于没有更新的网页不进行采集。

通过这样的方式可以提高信息采集的效率，在一定程度上减少由于股票数据高实时性的特点所带来的网络、存储等资源的要求。市面上比较成熟增量采集系统有：Univ 与 WebFountain。

同时网页上的股票数据还存在延迟高等问题，例如百度股票的实时数据往往延时在1~2分钟不等，明显不能满足用户对于股票交易数据预警的需要。因此本系统采集上海证券交易所(http://www.sse.com.cn/)的一手数据，通过这样的方式确保系统采集的信息的及时性与准确性。

1.3 研究的意义

本课题研究的意义在于通过采集股票交易数据及相关信息及时向用户预警、推送其所关注的信息。股票交易数据指在股票交易时产生的相关数据包括：当前股价、卖1~5 交易价格和挂单量、委比、成交量等。相关信息为股票所具有的财务数据包括：市盈率、市值、ROE等财务数据。以及公司公告。对于以上信息可由用户预设的条件在成功抓取后第一时间推送用户、进行预警。这可以大大减少用户在股票数据网站浏览信息的时间与经历。也可以做到没有疏漏，避免由于获取信息不及时或未能获取而导致的投资损失，由此股票信息预警系统的研究非常有必要，对于股票信息的采集、预警系统的反战具有较大的意义。

1.4

研究的目标与内容

本课题的目标就是通过使用网络爬虫在股票信息这个垂直领域中获取数据实现股票数据的采集。主要采取增量式网络爬虫信息抓取技术。本系统需要为用户提供股票信息的预警与推送，并展示处理后的股票交易数据。

本课题的研究主要是根据上述需求，设计解决方案完成系统的设计开发。具体如下所示：

1) 需求分析：确定系统的边界、所需实现的功能。

2) 概要设计：将需求分析得到的数据流图转化为软件结构图以及数据结构图。

3) 详细设计：对概要设计进行细化，详细设计各模块的算法实现与所需数据结构。

4) 系统实现：根据概要设计，使用C#语言进行实现，爬虫部分使用多线程、定时抓取的信息获取方式。

1.5 论文的组织安排

论文各章节安排如下：

第1章：绪论。对股票信息预警系统开发的背景与研究意义进行了介绍，同时介绍了本文的主要工作与结构安排。

第2章：相关理论与技术概述。对股票信息预警系统开发所需的相关理论与技术进行了简明概述，包括此次开发所需的软件开发环境、C#技术、爬虫技术、Html解析技术、Json格式化技术，以及Oracle 11g 数据库管理系统。本章为系统的开发提供了理论基础以及技术准备，是本文的理论与技术基础。

第3章：股票信息预警系统的需求分析。分别从市场、经济、技术等角度讨论了本系统的开发可行性。结合对于股票信息预警系统的需求，通过UML建模明确系系统开发的功能，明确开发技术要求。

第4章：股票信息预警系统的设计。完成系统结构设计、模块开发以及数据库设计。

第5章：股票信息预警系统的实现。根据需求分析与设计阶段的结果，确定系统开发方式。通过系统界面图、流程图以及关键代码对系统的重要模块的实现进行概述。

第6章：股票信息预警系统的测试。首先对于系统所需的基本方法、类进行必要的单元测试，选用黑盒测试方法。在之后的系统集成阶段对系统功能的实现进行集成测试。直至软件的效果达到需求分析中的要求。

第7章：总结与展望。对股票信息预警系统开发过程进行总结，并对去得结果进行分析。展望系统进一步的研究方向以及可以改进的部分。

2 股票信息预警系统的相关理论与技术概述

2.1 信息采集系统概述

信息采集系统指从非结构化的信息、或者有大量冗余、噪声的文件中将所需的信息抽取出来保存至关系型数据库中的软件系统。

对于数据源为网页的采集系统往往采用网络爬虫技术

2.2 网络爬虫概述

网络爬虫（Web Crawler）是指按照一定的规则，自动地抓取互联网信息的程序或者脚本。常见网络爬虫根据实现技术分类有通用(General Purpose)、增量(Incremental)、聚焦(Focused)、深层(Deep)等。在实际应用中往往需要将几类技术相互结合。

2.2.1 网络爬虫的工作流程

对于本程序由于股票的页面相对固定，因此可以采取将股票代码作为一个线性表，对每个股票代码进行遍历获取网页。另外还要对获取的信息与数据库中保存的信息进行比较，避免重复。

图2-1网络爬虫工作流程

爬虫不从URL集合中提取URL，按顺序访问获取URL所对应的HTML代码，这样就可以手机每个网页中的信息，直至集合中

的所有URL皆已被访问，爬虫即停止工作。

2.2.2 网络爬虫的框架

网络爬虫是信息采集系统中的核心，利用URL访问页面，从而获取网页中的信息、数据。

图2-2 网络爬虫的基本框架

说明：

* 调度模块：负责URL的调度，可以实现分布式爬虫的需求。
* 网页下载模块：负责个别行业URL下载HTML，可以设置请求方法、Cookie等。
* 页面处理模块：负责将HTML解析为所需的数据信息，一般传入为Html或Json 数据源。
* 数据保存模块：负责对所获取的数据信息存入相对应的数据库中。

2.2.3 增量爬虫

考虑到股票数据具有时效性的特点。因此需要平凡的对网页进行抓取。如果仅使用普通的爬虫将对系统所运行环境的网络、存储提出相当高的要求。因此本系统将采用增量爬虫。顾名思义就是仅抓取增量的网络信息。另一种解释是增量爬虫可以根据网页变化的频率进行数据采集工作，来获取新的网页数据并在同时将已抓取的数据进行更新。

增量式爬虫的特点可以总结为能快速的对抓取的数据进行更新，节省了不必要的重复执行，降低了对于运行环境的要求。大致的实现方法如下：

* 根据股票数据的刷新频率来决定爬虫的采集周期。
* 对于不同的数据采用不同的爬虫进行抓取。

3 需求分析

需求分析是软件工程项目开发的第一阶段，也是基础。该阶段通过对用户的需求进行分析，识别和确认出系统中的功能点，以及所需达到的目标，总结归纳得出系统的性能、功能需求。本章将对股票信息预警系统的需求分析展开详细的论述，讨论程序系统在当下经济、技术条件的开发可行性，以及系统的总题需求。同时根据用户需求，明确功能、技术要求。

3.1 系统开发可行性分析

该部分讨论在当前技术、经济条件下，是否能够带来相应的效益、是否能实现用户的需求。因此，可行性分析的实质就是明确开发系统的约束与限制条件。对股票信息预警系统的可行性分析从下面几个方面展开

3.1.1 市场可行性

本次开发的股票信息预警系统主要定位于独立投资者。随着经济的发展越来越多的独立投资人参与到了股票交易中。独立投资人由于其所具有的资金有限、股票操作难度较低，但迫于新的量化基金的诞生与互联网上的爆炸式的信息量，再采取原来的被动式接受信息方式必然会导致投资收益下降，乃至造成由于操作、信息而导致的亏损。而此次开发的股票信息预警正好迎合了独立投资人对于股票信息筛选、推送、预警的需求。因此，从市场角度考虑，本系统具备市场可行性。

3.1.2 经济可行性

目前，绝大多数的独立投资者都已经使用了股票交易软件进行交易。但由于股票交易软件的信息多、操作步骤复杂、可支持的预警形式单一等问题，极易给用户造成投资损失。而采用股票信息预警系统就可以有效的规避这一损失，提高综合股票投资收益。当然，本系统的部署与实施需要耗费相应的成本，但由于本软件可在普通个人电脑上运行且配置需求不高。综合成本与收益来看对于股本在10万元人民币及以上的用户所获得的收益将远大于成本。因此，从经济角度来看，本系统的开发具备经济可行性。

3.1.3 技术可行性

本次开发的股票信息预警系统是针对独立投资者的信息预警系统，采用C#为开发语言，C#是非常成熟的面向对象语言，是主流的Win from 程序开发技术。数据库采用Oracle 11g，具有非常强大的数据管理能力，通过其支持的PL/SQL 语言可以实现非常灵活、灵活的数据操作。因此本系统采用了当前主流的Client\Service 结构开发技术与数据库支持系统，可以很好的满足开发需求，且在未来也有较为强大的适应能力。因此，从技术可行性的角度考虑，本系统开发具备技术可行性。

3.1.4 操作可行性

随着信息化时代的发展以及国内经济的快速发展，个人电脑在国内迅速普及，而绝大多数的股票独立投资者都具有操作股票投资软件的经历。本系统采用 Microsoft Visual Studio 2017 为继承开发环境，该工具的 .NET framework 框架具有良好的 Win from 交互界面。同时本系统用户交互界面部分采取了与市面上常见的股票交易软件如“同花顺”等相类似的界面布局，操作简便，对于计算机水平较低的用户提供了充足的预设预警条件，可以极为便捷的设置预警。而对于具有较高计算机素养的用户也提供了具有半编程性的界面以支持复杂预警的设置。因此，从操作可行性考虑，本系统的开发具备操作可行性。

综上所述，此次股票信息预警系统迎合了独立投资者的需要，同时在经济、技术、操作方面都具备相当的开发可行性。由此可以得出本系统的开发是完全可行的结论。

3.2 系统功能需求分析

为了保证此次开发的股票信息预警系统符合独立投资者对于股票信息预警的需求，作者深入了解当下股票重要的数据、信息，以及对于获取信息展示、处理的要求。

3.2.1 数据要求

所需获取的股票交易及相关数据有：
* 股票历史交易数据包括：
* 日期
* 股票代码
* 当日开盘价格；
* 当日收盘价格；
* 当日最高价格；
* 当日最低价格；
* 当日成交金额。
* 股票日交易数据包括：
* 股票代码；
* 昨日收盘价格；
* 日期；
* 时间；
* 当时成交价格；
* 当时成交量。
* 股票快照数据包括：
* 时间；
* 买一~买五、卖一~卖五价格；
* 买一~买五、卖一~卖五挂单量；
* 当日最高；
* 当日最低；
* 当日开盘价格；
* 今日累计交易量；
* 今日累计交易金额；
* 当前竞买价格；
* 当前竞卖价格；
* 当前价格；
* 涨跌额；
* 涨跌百分比。
* 股票公告信息包括：
* 股票代码
* 公告标题
* 日期
* 所在页面的URL
* 股票财务数据信息包括：
* 股票代码；
* 时间；
* 收益；
* 市盈率(PE);
* 净资产；
* 市净率；
* 营收；
* 净利润；
* 毛利率；
* 净利率；
* ROE；
* 负债率；
* 总股本；
* 总市值；
* 每股未分配利润
对于股票日交易数据、快照数据要求实时更新，当有公告或股票财务数据发生变化时需迅速更新。

3.2.2 功能需求

通过对部分独立投资人的采访与调研，总结归纳出如下需求：

* 能及时获取数据要求部分的数据要求必须保证其准确性。

* 能够对抓取获得的数据进行有效的展示，其中的历史数据、日交易数据可以转化为对应的K线图及日行情折线图便于观察。

* 可以根据预设条件设置预警条件通过手机短信或邮件的形式进行预警推送。

* 能够支持定制化的预警需求。

3.2.3 总体需求

通过对上述需求的分析，我们可以分析出用户的需求。股票信息预警系统由于是面向个人投资者因此参与者仅为客户一个角色，用例图如下：

图3-1股票信息预警系统用例图

系统爬虫部分自动在交易日启动采集页面功能，按预设条件采集页面，然后存储，过滤重复数据、冗余数据，并及时更新抓取的数据。其中包含对于页面URL的解析，用于对URL进行处理。

页面解析功能是对采集到的页面进行结构化转换，将页面中的数据进行初步的结构化处理。

数据处理功能是指在页面解析之后根据传入的结构化数据进行整理、分类、比对并存储至关系型数据库中。

当数据被采集后，用户可以对股票信息进行浏览与预警设置。

用户通过对股票的实时快照如价格上破、价格下破，股票交易日数据中的累计成交额、历史数据中的10日均线等各类预设预警条件进行配置，以便用户可以在这些条件发生后尽快的通过预警系统获知这一情况。

3.2.4 数据爬虫功能需求

股票信息预警系统的根本是对于股票数据、信息的采集们需要将所需的与股票相关的网页采集、获取其中的数据、信息并存储起来。采集的网页主要包括上海证券交易所官网、东方财富网、新浪财经等主流股票信息网站。采集的内容如下图所示：

图3-2 爬虫数据采集用例图

采集目标网站主要以最具权威的上海证券交易所官网为对象，其次是新浪财经网，最后考虑如东方财富网之类的金融信息网站。最大程度的保证数据、信息的时效性、准确性。

页面采集后会根据页面中数据刷新时间这一信息进行判断，如果与已存入数据库中的值相一致，则放弃采集。

3.2.5 股票数据浏览模块需求

股票信息浏览可以帮助用户了解股票的很多信息，包括快照信息、财务数据、最新公告、历史数据、日交易数据。

用户首先选择需要浏览的股票集合。在本系统中主要分为全部浏览以及浏览关注列表两种。

在这两中列表浏览形式下用户可以快速了解股票的快照数据，并通过点击对应股票来浏览股票的其他更详细的信息。

这些信息包括股票的公告、日交易数据、以及K线数据，同时日交易数据、K线数据还支持对其中的原子数据进行查询与浏览。

具体的浏览关系如下图所示：

图3-3 股票信息浏览用例图

3.2.6 预警模块需求

用户首先根据自己的需要或选择预设条件进行预警条件设置，预警系统将由爬虫抓取并存入数据库的数据与用户预设的条件进行比对，当满足用户的预设条件后，按用户设置的预警推送方式将相关的信息推送给用户。具体如下图所示：

图3-4 预警模块用例图

3.3 非功能需求

除了上上述的功能需求之外，由于本系统为实时预警系统，而且本系统需要长时间运作因此对于系统的稳定性、高性能有特别的需求。

3.3.1 稳定性

股票信息预警系统的基础是对于股票信息的采集与处理，而作为一个需要长时间运作的采集系统最基本的需求就是系统具有较高的稳定性。本系统从数据获取、数据展示、预警设置、预警发送各个环节彼此关联紧密。其中任意环节发生问题都想将导致系统的不稳定。甚至可能导致由于数据抓取的错误而发送错误预警导致用户反而蒙受损失。

3.3.2 高性能

本系统需要抓取股票的快照数据，其刷新频率大致为5~10秒，同时还要抓取各股票的公告、财务数据。可以说每天的数据量是较为庞大的。如果性能得不到保障，则每个交易日的数据采集作业将不可持续。这对于一个股票信息预警系统而言是极为危险的。

3.4 本章小结

需求分析是软件开发的第一个阶段、也是最为总要的阶段。通过对独立投资人的访谈与咨询确定了股票信息预警系统的功能需求。本章首先结合了系统的总体需求、数据需求、功能需求等进行了概要的描述。接着对爬虫模块、浏览模块、预警模块等各个功能的需求进行了详尽的描述，指出个模块所需完成的目标。最后提出了对系统所应具有的非功能性需求。

4 股票信息预警系统的设计

完成股票信息预警系统的需求分析后，根据软件工程的开发模式瀑布模型进入系统的设计阶段，系统设计阶段主要完成对系统的结构设计、系统模块设计与数据库表结构设计。本章将在确定的系统设计原则的原则上，对以上三个部分展开设计

4.1 系统结构设计

4.1.1 系统体系结构设计

根据需求分析对于系统的需求，以及当前软件体系结构的发展趋势，本次股票信息预警系统的开发将采用Client/Server（“简称B/S”）体系结构进行设计。由于系统中的爬虫部分、预警部分需要一直开启处在运行状态，要求运行的机器不能关闭，同时会占据较多的系统资源。因此这两部分的子系统放置在服务器端进行计算运行，而将对系统要求较低的数据浏览、预警设置等功能放置在客户端运行。本系统的体系结构如图所示：

图 4-1 股票预警系统结构图

4.1.2 系统物理结构设计

为了满足系统的高性能与高稳定性的需求，本系统由应用服务器、数据库服务器、客户机组成。

| 指　标 |
| --- |
| 疑似剽窃观点 |
| 1. 因此，从经济角度来看，本系统的开发具备经济可行性。 |
| 疑似剽窃文字表述 |
| 1. 因此，从技术可行性的角度考虑，本系统开发具备技术可行性。<br>3.1.4 操作可行性 |

## 2. 基于网络爬虫的股票信息预警系统的设计与实现_第2部分

总字数：11876

相似文献列表　文字复制比：5.1%(605)　疑似剽窃观点：(0)

| | | |
| --- | --- | --- |
| 1 | 珠宝管理销售系统软件建模与分析.doc 20页-高清全文免费预览-max文档<br>-《互联网文档资源（http://max.book118.c）》- 2015 | 1.6%（190）<br>是否引证：否 |
| 2 | 基于JAVA EE的服装电子商务平台开发与实现<br>谭阳 -《大学生论文联合比对库》- 2016 | 1.3%（160）<br>是否引证：否 |
| 3 | 基于模块化方法的住宅精装修施工工艺规划系统的研究<br>张昕(导师：唐任仲) -《浙江大学硕士论文》- 2014 | 1.3%（157）<br>是否引证：否 |
| 4 | 基于WAT的区域卫生信息平台的架构与实现<br>白昌盛(导师：党伟超) -《太原科技大学硕士论文》- 2014 | 1.3%（150）<br>是否引证：否 |
| 5 | 无线电管理部门监测数据资源共享平台设计与实现<br>肖储宁(导师：许智宏) -《河北工业大学硕士论文》- 2014 | 1.2%（138）<br>是否引证：否 |
| 6 | 基于Android的经管学院请假管理APP设计与实现<br>罗兴 -《大学生论文联合比对库》- 2016 | 1.1%（135）<br>是否引证：否 |
| 7 | 1203100224-荣胜华-计算机科学与技术<br>荣胜华 -《大学生论文联合比对库》- 2016 | 1.1%（129）<br>是否引证：否 |
| 8 | 基于B/S的期货行情分析系统的设计和实现<br>隋春霞(导师：李大奎) -《大连理工大学硕士论文》- 2014 | 1.0%（120）<br>是否引证：否 |
| 9 | Oracle Developer - David Dai -- Focus on Oracle - 博客频道 - CSDN.NET<br>-《网络（http://blog.csdn.net）》- 2012 | 0.9%（112）<br>是否引证：否 |
| 10 | Oracle Developer - David Dai -- Focus on Oracle - 博客频道 - CSDN.NET<br>-《网络（http://blog.csdn.net）》- 2012 | 0.9%（112）<br>是否引证：否 |
| 11 | 网上书店系统的设计与实现毕业论文 - 豆丁网<br>-《互联网文档资源（http://www.docin.com）》- 2012 | 0.8%（100）<br>是否引证：否 |
| 12 | 科研团队信息管理系统数据建模分析与实现<br>倪良智(导师：赵姝) -《安徽大学硕士论文》- 2013 | 0.8%（98）<br>是否引证：否 |
| 13 | 面向服务架构的EAM系统研究与设计 | 0.8%（94） |

| | | |
|---|---|---|
| | 秦虹(导师：骆德渊;徐济高) - 《电子科技大学硕士论文》- 2011 | 是否引证：否 |
| 14 | PowerDesigner 物理数据模型（PDM） 说明 - David Dai -- Focus on Oracle - 博客频道 - CSDN.NET | 0.8%（94） |
| | -《网络（http://blog.csdn.net）》- 2012 | 是否引证：否 |
| 15 | 山西交警办公自动化系统的设计与实现 | 0.8%（94） |
| | 卢红波(导师：朴勇) - 《大连理工大学硕士论文》- 2014 | 是否引证：否 |
| 16 | 变电站综合自动化系统监控软件的需求分析与设计 | 0.8%（93） |
| | 武常刚(导师：程保中) - 《北京邮电大学硕士论文》- 2009 | 是否引证：否 |
| 17 | 沪江博客 - 逆风飞飏 - 软件测试面试题目（三） | 0.8%（93） |
| | -《网络（http://blog.hjenglis）》- 2012 | 是否引证：否 |
| 18 | 基于ASP的财政办公自动化系统的设计与实现 | 0.8%（91） |
| | 黄旭辉; -《企业家天地》- 2008 | 是否引证：否 |
| 19 | 基于ASP的财政办公自动化系统的设计与实现 | 0.8%（91） |
| | 徐刚;杨林海;李绍伟; -《OA'2010办公自动化国际学术研讨会论文集》- 2010 | 是否引证：否 |
| 20 | 基层医院绩效薪酬管理信息化平台的构建及实施 | 0.7%（87） |
| | 梁大伟;谭剑;黄煌镜;廖生武; -《信息化建设》- 2016 | 是否引证：否 |
| 21 | 基于net的人力资源系统的设计和实现 | 0.7%（83） |
| | 孙大伟(导师：王静;崔宏伟) - 《电子科技大学硕士论文》- 2010 | 是否引证：否 |
| 22 | 药典系统的设计与实现 | 0.7%（83） |
| | 黄悦(导师：齐德昱;黄星光) - 《华南理工大学硕士论文》- 2015 | 是否引证：否 |
| 23 | 山东省植物信息系统的设计与实现 | 0.7%（82） |
| | 马玉强(导师：单世民) - 《大连理工大学硕士论文》- 2009 | 是否引证：否 |
| 24 | ERwin在教学管理系统设计中的应用研究 | 0.6%（76） |
| | 邵莉;李清茂; -《攀枝花学院学报》- 2010 | 是否引证：否 |
| 25 | 基于SSH框架的可在线支付的网上书店系统设计与实现 | 0.6%（69） |
| | 李洋 -《大学生论文联合比对库》- 2016 | 是否引证：否 |
| 26 | PowerDesigner在航空军事运输动态监控系统设计中的应用 | 0.6%（68） |
| | 苑德春;马忠俊;张远大;葛同民; -《军事交通学院学报》- 2009 | 是否引证：否 |
| 27 | 基于组件技术的电信资源管理系统的研究和实现 | 0.6%（68） |
| | 刘敏(导师：曾志民) - 《北京邮电大学硕士论文》- 2006 | 是否引证：否 |
| 28 | 基于树比较的Web页面主题信息抽取 | 0.4%（44） |
| | 朱梦麟;李光耀;周毅敏; -《微型机与应用》- 2011 | 是否引证：否 |
| 29 | 基于MySQL的无线抄表供暖数据库系统的设计与优化 | 0.3%（32） |
| | 单菲(导师：陈冬岩) - 《山东大学硕士论文》- 2014 | 是否引证：否 |
| 30 | 光功率指标在传送网管理中的应用研究与实现 | 0.3%（30） |
| | 黄婷婷(导师：亓峰) - 《北京邮电大学硕士论文》- 2010 | 是否引证：否 |

原文内容

物理结构如下图所示：

图4-2 股票信息预警系统物理结构

若服务器性能较好、其中的爬虫服务器与预警服务器可以采用同一个。

4.1.3 功能模块划分

通过对系统需求的研究与分析，对系统架构按功能模块划分如下图：

图 4-3 股票信息预警系统模块架构图

从图中可以看出股票信息预警系统的架构设计，可以分为五个模块：爬虫模块、网页解析模块、数据处理、浏览模块、预警模块。系统从采集数据存入数据库中作为临时文件、将这些临时文件从数据库中取出交由网页解析模块进行解析，通过数据处理将网页中需要的目标信息进行抽取、整理存入数据库中。预警模块与浏览模块根据数据库中处理过的数据分别进行信息的预警与信息的展示。

4.2 股票信息预警系统类设计

股票信息预警系统的总体类图设计如下图所示：

图4-4 股票信息预警系统类图

通过调用统一的动态链接库简化了代码的调用层级关系，实现了软件设计中的低耦合、高内聚的要求。

4.2.1 系统动态链接库类图

图4-5 股票数据类图

这一部分是根据数据需求对抓取的股票数据、信息进行面向对象化建模，每一个数据库中所存储的数据都有对应的模型。具体对应关系如下：

* ECNOData：为股票的财务数据；
* StockLineData：为股票日数据；
* StockHisData：为股票历史数据；
* StockSuggest：为输入提示框所需的数据；
* StockList：股票列表（URL集合）；
* Announcement：为股票公告信息；
* SnapData：为股票快照数据，可具体细分为：
* Data：股票的快照数据具体参照第三章《数据需求》；
* Gopicture：为网站提供的交易数据可视化图
* Dapandata：大盘交易数据；

其中StockSuggest、StockLineData、StockHisData都调用了BaseCrawler类，BaseCrawler为针对本系统所改进的增量爬虫。该增量爬虫的设计如下图：

图4-5 增量爬虫设计类图

下面对上图中的类进行说明。HttpItem为爬虫Http。 HttpResult为Http返回参数类， HttpHelper为爬虫主体代码。

图4-6 数据库连接类图

这一部分是爬虫子系统、预警子系统、客户端程序与数据库进行交互的共通类、由于不同的C#连接Oracle数据库方式有不同的优势因此这里同时给出三种链接方式对于不同的需求采用不同的连接方式。

4-7 Json格式化类图

由于各网站的通信数据不一定都是标准的Json格式。例如上海证券交易所的日数据返回字符串：

“({"code":"600004","total":3354,"begin":3353,"end":3354,"kline":[[20170428,16.01,16.01,15.79,15.95,7455272]]})”

缺少必要的Json标签。正确的Json字符串应为：

"{"code":"600000","total":"4105","begin":"0","end":"1","kline":[{"date":"19991110","open":"29.50","high":"29.80","low":"27.00","close":"27.75","amount":"174085000"}]}"

因此需要通过专门的过程对获取的字符串进行Json格式化。

图4-8 预警子系统类图

这一部分是预警子系统的预警推送部分类图，用户根据不同信息预警的重要程度，采用短信（SMS）或电子邮件的形式对预警的信息进行推送，由于短信的推送延迟要小于电子邮件因此，对于较为重要的股票信息预警应当采用短信的推送方式。

以上就是本系统主要的类图，通过对功能需求有效的拆分、生成对应的类，有效的提高了软件代码的复用，降低了软件维护、升级、二次开发的成本。

4.3 股票信息预警系统顺序图设计

4.3.1 爬虫子系统顺序图设计

股票信息爬虫采集系统的数据主线为爬虫系统从互联网页面抓取所需的股票页面、进行解析、处理、最终存储至数据库。如下图所示：

图4-9 爬虫子系统采集数据顺序图

爬虫子系统首先通过数据采集模块向互联网发送请求获取要抓去的页面，然后保存页面进行格式化处理，之后提取格式化数据中的信息，最后存入数据中，之后由数据库查询下一个需要爬取的URL传递至数据采集模块，进行循环。

4.3.2 预警子系统顺序图

图4-10 预警子系统预警顺序图

预警子系统通过查询获取数据库中股票信息、数据以及用户预设的预警条件，将两者进行比较当预设条件被满足时向用户推送预警结果。

4.4 数据库设计

由于系统进行数据处理的数量十分巨大，因此为了满足系统的需要，系统采用Oracle数据库。同时为了降低数据库压力，数据库中只保留最小数据集，即仅存储元数据、所有可以通过公式计算出的数值都将在预警程序运行时通过计算获得，同时这样也可以减少数据库传输数据的字节数，进一步提高性能。

4.4.1 数据库需求分析

除了采用性能较好的数据库，数据库结构设计的优劣也对数据库性能起到了至关重要的影响。良好的数据库结构在保证了数据完整、一致的同时，还能进一步提高数据存取的效率。同时，合理的数据库设计也能在一定程度上降低系统开发的难度。

4.4.2 数据库概念结构设计

概念设计是根据特定的方法将现实世界构建为一个不依赖于具体实现的数据模块，即概念模型（Conceptual Data Model）。概念模型独立于机器，是抽象级别最高的数据模型。最常见的概念模型作图法是实体-联系模型，亦称E-R图。E-R图由实体、属性、关系三者构成。本系统设计众多实体、属性，由于ER图所占篇幅较长，因而改用另一种概念模型作图法IDEF1X图。IDEF1X是IDEF系列方法中IDEF1的扩展版本,是在E-R(实体联系)法的原则基础上,增加了一些规则, 使语义更为丰富的一种方法。用于建立系统信息模型。本系统概念模型图如下：

图4-11 股票信息预警系统概念模型图（CDM）

4.4.3 数据库物理模型

物理数据模型（Physical Data Model），提供了系统初始设计所需要的基础元素，以及相关元素之间的关系。使用物理数据模型，可以在系统层实现数据库。数据库的物理设计阶段必须在此基础上进行详细的后台设计，包括数据库的存储过程、操作、触发、视图和索引表等，将上文中的CDM模型进行处理生成如下PDM图：

图4-12 股票信息预警系统物理模型（PDM）

通过上文的CDM与PDM模型构建数据库结构创建如下表：

表4-1 ANNOUNCEMENT表结构

名称说明数据类型长度主键外来键

CODE 股票代码 VARCHAR2(20) 20 TRUE TRUE

URL 公告URL VARCHAR2(500) 500 TRUE FALSE

TITLE 标题 NVARCHAR2(200) 200 FALSE FALSE

DAYS 日期 DATEFALSE FALSE

ALARMED 是否已预警 VARCHAR2(20) 20 FALSE FALSE

表4-2 BASE_INFOR表结构

名称说明数据类型长度主键外键

PHONE 手机号码 VARCHAR2(20) 20 FALSE FALSE

EMAIL 邮箱 VARCHAR2(200) 200 FALSE FALSE

STAMP_TAX 印花税 NUMBERFALSE FALSE

SEC_CHARGES 交易手续费 NUMBERFALSE FALSE

STOCK_TRADING_FEES 交易税 NUMBERFALSE FALSE

TRANSFER_FEES 过户费 NUMBERFALSE FALSE

BROKERAGE 佣金 NUMBERFALSE FALSE

表4-3 BASE_INFOR表结构

名称说明数据类型长度主键外来键

SET_NAME 组名 VARCHAR2(200) 200 TRUE FALSE

ID 编号 NUMBERTRUE TRUE

表4-4 FOCUS_LIST表结构

名称说明数据类型长度主键外来键

CODE 股票代码 VARCHAR2(20) 20 TRUE TRUE

表4-5 STOCK_ECNO_DATA表结构

名称说明数据类型长度主键外来键

CODE 股票代码 VARCHAR2(20) 20 TRUE TRUE

DAYS 日期 DATETRUE FALSE

INCOME 净收益 NUMBERFALSE FALSE

PE 市盈率 NUMBERFALSE FALSE

BVPS 净值 NUMBERFALSE FALSE

PB 市净率 NUMBERFALSE FALSE

续表STOCK_ECNO_DATA

REVENUE 营收 NUMBERFALSE FALSE

REVENUEYOY 营收同比 NUMBERFALSE FALSE

NETPROFIT 净利润 NUMBERFALSE FALSE

NETPROFITYOY 净利润同比 NUMBERFALSE FALSE

GROSSMARGIN 毛利率 NUMBERFALSE FALSE

NETMARGIN 净利率 NUMBERFALSE FALSE
ROE ROE NUMBERFALSE FALSE
DEBTRATION 负载率 NUMBERFALSE FALSE
EQUITY 总股本 NUMBERFALSE FALSE
VALUE 总值 NUMBERFALSE FALSE
UDPPS 每股未分配利润 NUMBERFALSE FALSE
表4-6 STOCK_HIS_DATA表结构
名称说明数据类型长度主键外来键
CODE 股票代码 VARCHAR2(20) 20 TRUE TRUE
DAYS 日期 VARCHAR2(50) 50 TRUE FALSE
OPEN 开盘 NUMBERFALSE FALSE
HIGH 最高 NUMBERFALSE FALSE
LOW 最低 NUMBERFALSE FALSE
CLOSE 收盘 NUMBERFALSE FALSE
AMOUNT 成交总量 NUMBERFALSE FALSE
表4-7 STOCK_LINE_DATA表结构
名称说明数据类型长度主键外来键
CODE 股票代码 VARCHAR2(20) 20 TRUE TRUE
DAYS 日期 VARCHAR2(20) 20 TRUE FALSE
TIME 时间 VARCHAR2(20) 20 TRUE FALSE
PRICE 价格 NUMBERFALSE FALSE
VOLUME 成交量 NUMBERFALSE FALSE
表4-8 STOCK_LIST表结构
名称说明数据类型长度主键外来键
CODE 股票代码 VARCHAR2(20) 20 TRUE FALSE
NAME 股票名称 VARCHAR2(50) 50 FALSE FALSE
LAUNCH_DATE 上市日期 DATEFALSE FALSE
GENERAL_CAPITAL 流通股本 NUMBERFALSE FALSE
NEGOTIABLE_CAPITAL 总股本 NUMBERFALSE FALSE
表4-9 STOCK_WARNING表结构
名称说明数据类型长度主键外来键
CODE 股票代码 VARCHAR2(20) 20 TRUE TRUE
NAME 预警名 VARCHAR2(20) 20 TRUE FALSE
STATE 状态 VARCHAR2(20) 20 FALSE FALSE
EXPLAIN 说明 VARCHAR2(200) 200 FALSE FALSE
WAYS 方式 VARCHAR2(20) 20 FALSE FALSE
LOGISTICS 逻辑 VARCHAR2(2000) 2000 FALSE FALSE
表4-10 STOCKSUGGEST表结构
名称说明数据类型长度主键外来键
CODE CODE VARCHAR2(20) 20 TRUE TRUE
NAME NAME VARCHAR2(50) 50 FALSE FALSE
PY PY VARCHAR2(20) 20 FALSE FALSE
EXPLAIN 说明 VARCHAR2(200) 200 FALSE FALSE
WAYS 方式 VARCHAR2(20) 20 FALSE FALSE
LOGISTICS 逻辑 VARCHAR2(2000) 2000 FALSE FALSE
表4-11 WARNING_QULAFICATION表结构
名称说明数据类型长度主键外来键
ID 编号 NUMBERTRUE FALSE
SHOW_NAME 显示名称 VARCHAR2(200) 200 FALSE FALSE
QUALIFICATION_CODE 条件代码 VARCHAR2(200) 200 FALSE FALSE
TARGET_NUM 目标数字 NUMBERFALSE FALSE
SIGN 符号 VARCHAR2(2) 2 FALSE FALSE

4.5 详细设计

概要设计完成后，就需要进行系统的详细设计。在详细设计阶段主要是完善系统各个模块的流程与类。

4.5.1 爬虫模块

爬虫模块是股票信息预警系统的第一个模块，主要是用于对互联网上股票页面进行抓取、获取股票信息、数据等一系列网页。

根据系统对于可靠性、时效性的要求，抓取上海证券交易所的官方网站（http://www.sse.com.cn/ 一下简称上证官网）是最佳的目标。由于是第一手数据源所以在数据延迟上是全网最小的。同时由于是官方网站页面比较简洁，其他网站往往需要加载大量无效的图片、flash文件延缓了爬虫抓取页面的速度。同时由于页面简洁也为之后页面元素的解析带来的遍历。

但上证官网也有其缺陷，那就是信息纬度较少，部分例如股票财务数据、股票公告的发布都不能从该网站抓取到。

4.5.2 爬虫模块活动图

爬虫采集模块的活动图如下图所示：

图4-13 爬虫采集模块的活动图

从图中可以发现爬虫模块先初始化一个股票信息URL队列；之后获取队首元素URL并将其从队首移去，进行页面的抓取；之后重复这个过程就可以获取所有代码的信息。由于股票的代码、数量相对固定，因此不必在日常的抓取中更新URL队列。

4.5.3 爬虫模块顺序图

爬虫采集对象中有不同的网页需要采集、根据页面的类型不同应采用不同的抓取策略，爬虫模块内部执行一次抓取流程的顺序图如下：

图4-14 爬虫模块顺序图

如图所示为爬虫抓取股票公告信息的过程。Announcement 收到获取公告的请求初始化获取股票队列；取出队首元素与类的属性URL进行拼接向HtmlDocument发出请求获取数据；HtmlDocument 获取数据后将数据传回给Announcement类创建实例，之后调用Announcement的SaveToDB()方法将实例结构化存入数据库中，数据库在获取Announcement实例的数据后通过数据库的主键的设置确保该数据是否已经存入数据库保证不重复插入数据。之后重复这一过程直至URL队列中的所有元素都已经被使用。

4.5.4 网页解析模块

本系统中网页解析主要分两类：

* 对类Json字符串的动态加载部分进行Json序列化处理；

* 对Html页面采用XPath技术解析获取目标标签中的数据；

对于第一种情况的主要处理方式是通过网页调试工具观察返回字符串与Json标准格式的区别，或者观察网页中的Js代码是如何处理返回的类Json字符串的。但由于网页常常会进行Js混淆、紧凑化等处理手段导致Js代码难以阅读、理解，因此主要是第一种方法进行处理。第一种方法常用的技术手段有正则表达式匹配替换以及字符串遍体替换这两个方法。本系统采用字符串遍历替换、重新构造的方法，将类Json字符串序列化。

对于第二种情况采用XPath技术。XPath即为XML路径语言，它是一种用来确定XML（标准通用标记语言的子集）文档中某标记位置的语言。XPath基于XML的树状结构，提供在数据结构树中找寻节点的能力。由于Html在大多数情况下具有与XML相类似的树状结构。因此往往可以使用这一技术来获取Html页面中所需要的数据。

4.6 本章小结

本章针对股票信息预警系统的设计过程进行了详细的阐述。通过对需求分析结果的研究，进行系统物理结构、软件结构的设计，采用面向对象的软件开发技术，利用UML分析、设计系统。最后对数据库的概念模型以及物理模型展开了详细设计，并根据模型完成了数据库表结构设计，为系统的后续实现夯实了基础。

5 股票信息预警系统的实现

在前两个章节中，明确了系统的需求，以及设计方法。本章将按前文所确定的要求、设计，利用软件开发工具、技术完成股票信息预警系统的实现。

5.1 系统实现的环境

5.1.1 系统实现的软件环境

本系统的开发采用的IDE为Visual Studio 2017 ,数据库为 Oracle 11g，主要开发语言为C#，.Net 框架版本为 .NET Framework 4.5.2 。由于本系统采用C#的Win From 应用程序开发模版，因此必须在Windows XP及以后操作系统上使用。本次系统的数据库以及爬虫模块、预警模块都安装在服务器上，因此对于客户端的要求是可以安装运行客户端程序。

5.2 系统模块的实现

本程序主要涉及爬虫模块、预警模块、数据浏览模块，各个模块之间存在众多功能，以下仅对这三个模块的主要功能的实现过程进行介绍。

5.2.1 爬虫模块的实现

根据系统设计阶段的要求，爬虫模块应该实现对不同的网页都有抓取能力，可以根据不同的数据刷新周期，设计不同的爬虫抓取数据周期。应可以获取包括股票快照数据、日交易数据、历史交易数据、财务数据以及股票发布公告的功能，并将获取

的相关数据存入数据库中。

具体实现步骤如下首先创建三个基础类 HttpHelper用于发起http请求获取返回结果，HttpItem用于组织发起http请求时的头部数据包括使用Get\Post方法、设置Host信息、设置Encoding 字符集等信息。HttpResult为Http请求的返回结果包括Cookie，html代码、重定向URL等。通过使用这三个基类对.Net 自带包封装后，可以实现抓取时无视编码、证书、Cookie等网页问题，可以同时使用Get和Post请求， 简化了对于Cookie，证书，代理的设置。C#常遇到的网页编码问题也通过类中的SetEncoding(),GetByte()方法解决大大简化了之后开发具体爬虫的难度与代码量。

之后利用将这三个基类用BaseCrawler类进行封装进一步简化代码复杂度，使对爬虫的调用简化为BaseCrawler().Run(URL)，返回值即为根据URL获取的html代码。

至此爬虫模块获取页面信息就完成了。之后就是对与Html代码的解析与格式化存入数据库。

根据第四章数据库表设计建立对应的面向对象模型。下文以StockHisData（股票历史数据）。

首先根据数据表4-6 STOCK_HIS_DATA进行面线对象设计，代码如下：

图 5-1 股票历史数据建模代码图

通过一下代码将获取的Json 字符串转化为对象。

图 5-2 股票历史数据实例填充数据代码图

之后将历史数据实例转化为DataTable格式的结构化数据，最终存入数据库中。

图 5-3 股票历史数据实例转化为DataTable数据代码图

这样就完成了从URL到抓取数据进行处理、到转化为结构化数据存入数据库的过程。

5.2.2 预警模块的实现

根据系统设计阶段的要求，预警模块应可以通过将用户预设的条件与数据库中的数据进行比较，在用户设置数据被满足时，通过短信、邮件的形式进行提醒。

实现方法如下，将用户预设的条件以字符串的形式存入数据库。例如用户创建了一个名为“快买入”的预警条件为在股票‘600000’当前价格上破 10.8 元并且交易量超过1000手时通过短信的方式进行预警。那么在数据库中表STOCK_WARNING中将存入这样的一条记录：

图 5-4设置预警后执行的数据插入语句图

当进入股票交易时间预警系统将自动启用对数据表STOCK_WARNING中state为“running”的数据进行测试比较。通过将数据提取至C#程序并重新组成新的SQL 语句并执行，如本例中该重新组成的SQL语句为：

图 5-5预警测试执行语句图

那么当条件被满足时返回值为1，反之则为负。预警系统将根据用于选择的预警方式从BASE_INFOR 表获取对应的手机号或邮箱地址进行预警的推送。

预警的效果如下图：

图 5-6通过邮件方式预警图

图 5-7通过短信方式预警图

至此预警模块的主要功能都基本完成了。

5.2.3 数据浏览模块的实现

根据系统设计阶段的要求，数据浏览模块应当可以对爬虫抓取并存入数据库的数据进行浏览，其中日交易数据与历史交易数据（K线）应具备可视化功能。数据展示模块主界面如下图所示：

5.3 股票信息预警系统的实现

图 5-8数据浏览主界面图

可对所有股票的实时交易数据进行浏览，刷新频率为5秒。支持翻页以及指定页面跳转添加了对于跳转目标页面页数的合理性验证。

图 5-9数据浏览主界面翻页图

同时支持对关注列表的股票进行浏览。

图 5-10对关注列表浏览图

双击列表中的股票就可以进入股票详细界面如下图，包括由股票日数据生成的折线图、股票的近3个公告、以及股票快照数据、财务数据，同时将鼠标在图像处滑动还可以看具体日数据。

图 5-11个股详细界面图

图 5-12左侧详细数据根据鼠标所指位置不同显示数据不同图

在这个界面同时可以通过点击收藏、取消收藏按钮将个股添加、移出关注列表。点击底部最新公告可直接打开浏览器并跳转至相关公告的新闻页面。

图 5-13点击公告连接后显示的公告页面图

点击查看K线可以浏览该股K线图，效果如下图所示：

图 5-14个股K线页面图

K线支持修改加载天数，同时显示区域也会跟随数据而实时变化。

图 5-15 K线加载数据变化图

5.4 本章小结

本章对于股票信息预警系统的部分模块实现做了简要论述。首先介绍了系统实现所需的软、硬件环境；通过流程图、系统界面以及部分关键代码对系统功能模块的运作方式进行了介绍。为了保障系统的可靠性、稳定性，还需要经过一系列的测试才能正式投入使用。因此，在下一张将对股票预警系统的测试进行介绍。

6 股票信息预警系统的测试

系统测试对于任何软件开发项目都是重中之重。爬虫模块是进行对网页采集，通过测试需要确认网页是否被正确抓取。预警模块是对用户预设条件进行比对来推送预警信息，需要通过测试确认预警是否正确推送。数据解析部分是对抓取获得的网页进行解析将数据存入数据库，通过数据库可以查看所解析数据是否正确。

6.1 单元测试

本系统采用将所有基类打包，形成动态链接库的形式，因此如果要进行单元测试只需要对动态链接库中的类进行测试即可。利用Visual Studio的创建单元测是功能对动态链接库中所有public访问级别的函数进行了测试。单元测试（模块测试）是开发者编写的一小段代码，用于检验被测代码的一个很小的、很明确的功能是否正确。通常而言，一个单元测试是用于判断某个特定条件（或者场景）下某个特定函数的行为。测试结果如下：

图 6-1线加载数据变化图

可见所有测试都通过，这表明所有代码都能按设计运行获取正确的结果与返回值。根据Visual Studio计算的语句覆盖率为85%。其中剩余未覆盖区域主要为try – catch 代码段。可以说程序进行了有效的单元测试。

6.2 测试用例

6.2.1 数据浏览主页面页码跳转功能测试

表6-1 数据浏览主页面页码跳转测试用例

测试项操作步骤预期结果实际结果比较缺陷严重程度

跳转输入正确的页码跳转至指定页面跳转至指定页面通过无发现

输入英文字符提示错误提示错误通过无发现

输入负数提示错误提示错误通过无发现

输入页码大于总页数跳转至最后一页跳转至最后一页通过无发现

图6-1线加载数据变化图

6.2.2 基本信息设置功能测试

表6-2 基本信息设置功能测试用例

测试项操作步骤预期结果实际结果比较缺陷严重程度

跳转输入正确的电话及邮箱保存修改成功保存修改成功通过无发现

错误的手机号提示错误提示错误通过无发现

错误的邮箱提示错误提示错误通过无发现

两者都错误提示错误提示错误通过无发现

图6-2确的手机号码更新成功

图6-3 错误手机号提示错误

图6-4 错误邮箱提示错误

图6-5 错误邮箱与手机号提示错误

6.3 测试结果

股票信息预警系统测试在代码级采用单元测试验证了各程序单元的正确性。同时对数据浏览界面采用黑盒测试法，达到了需求分析阶段的要求、可以上线运行。

此外，为了更加全面的对本系统的运行性能与功能的完整性进行进一步的测试。邀请了2位同学参与软件程序的试用，进过15天的试运行4月15日~5月1日。没有发现重大系统Bug。进过这次试运行，基本可以证明系统达到了需求分析中所确定的功能需求与技术需求。

6.4 本章小结

软件测试在整个软件开发流程中极为重要。本章对股票信息预警系统进行了单元测试与黑盒测试，根据测试的结果，确定了系统达到了需求分析阶段所确定的功能需求与技术要求。

7 总结与展望

7.1 总结

本文通过对基于网络爬虫的股票信息预警系统进行了研究，并最终实现了股票信息预警系统的主要功能，重点研究了股票数据、信息爬虫抓取模块、网页解析模块的爬虫架构。设计了对股票数据的留言模块，最后实现了由用户设置条件的股票预警模块。

本文通过对一般独立投资人的访谈与采访，获取了投资者对股票信息预警系统的需求，进行了需求分析明确了需要抓取股票数据、信息。根据最终数据存入数据库的需求，明确了系统所需的其他功能，包括网页解析、网页格式化处理、数据展示、股票预警等模块。

对系统进行设计时，将系统按模块分解分为爬虫模块、预警模块、浏览模块、解析模块、数据处理模块进行设计。由于系统的实时性要求导致的星系量巨大，因此将重点放在了类图设计和数据库设计，使得数据库能够完全覆盖常用股票数据、信息。

系统采用C#进行实现，采用增量爬虫应对一直增长的股票数据量。网页解析模块采用了XPath 以及正则表达式两种方法。数据展示模块使用户可以从不同的角度了解到每一只股票。

系统采集数据主要来自上海证券交易所可以保证采集信息的实时性、准确性。

通过本次系统开发，详细地了解了C#爬虫的编写方式，尤其是还接触了很多股票的相关知识，对于日后自己参与股票投资打下了坚实的基础。

7.2 展望

股票信息预警系统的基本功能都得到了实现，但随着对股票交易策略的深入、系统所需采集、解析的信息量会越来越大，比如近来越来越多人将社交软件数据用于股票交易上。这毫无疑问会对系统提出更高的要求，因此还需继续研究。

同时国内组件刮起了一阵量化交易之风、本股票信息预警系统所抓取、保存的数据还能够作为量化交易，回测的基本数据，因此本系统在量化交易领域能起到的作用也让人非常的期待。

致谢

首先我要感谢闫昱老师对我的悉心指导，闫老师平时工作时间非常紧张，但她总能抽出时间对我和其他同学的问题一一耐心解答。在整个毕业设计过程中一直定时督促、监督我系统的完成情况。并在这过程中指出我设计、想法上的错误，让我少走了很多弯路，就像一盏明灯一样为我照亮前行的道路。在此我不禁有感而发想对闫老师说一句谢谢您。

在此除了感谢我的指导老师闫昱老师，我还有感谢几位为我的毕业设计费心费力的同学、朋友。首先我要感谢蔡郁敏同学在我因为英语翻译遇到困难时的无私的帮助。其次我要感谢韩嘉凌同学多次和我沟通、告知我常用的股票信息，及稳定、实时性强的数据源。然后我要感谢学校的张荣鉴老师为我提供服务器与场地支持了我毕业设计的开发。

最后我要感谢我的母校—上海第二工业大学，在这四年中，学校为我们创造了大量的机会增进我们的专业能力无论是竞赛性质的蓝桥杯还是校三小项目无一不使我获益良多。正是学校对于我的悉心培养，让我做好了充足的准备走出象牙塔，进入社会，向着人生新的阶段发起挑战。

参考文献

[1]Dai S, Wu X, Pei M, et al. Big data framework for quantitative trading system[J]. Journal of Shanghai Jiaotong University, 2017, 22(2):193-197.

[2]Kumar N, Singh M. Framework for Distributed Semantic Web Crawler[C]// International Conference on Computational Intelligence and Communication Networks. IEEE, 2016:1403-1407.

[3]胡婧, 叶建木. 基于微博信息的股票交易预测研究[J]. 财政监督, 2017(5):108-111.

[4]张伟丽, 赵美珍. 政策不确定性、价格时滞与股票收益[J]. 经济研究导刊, 2017(6):58-60.

[5]苏若凡. 基于网络爬虫的股票信息预警系统的研究与实现[J]. 电子世界, 2014(16):124-124.

[6]卜永忠. 面向金融信息的主题爬虫研究与应用[D]. 哈尔滨工业大学, 2008.

[7]齐文龙. 基于爬虫技术的基金信息采集系统的设计与实现[D]. 天津大学, 2012.

[8]陈亮华. 基于网络爬虫的基金信息抽取与分析平台[D]. 华南理工大学, 2010.

[9]温玲. 基于爬虫技术的股价分析系统的设计与实现[D]. 北京大学, 2014.

[10]维克托.斯波朗迪.专业投机原理机械工业出版社.2010

[11]丁士峰等.C#典型模块项目实战大全电子工业出版社.2012

[12]王小科等.C#项目开发案例全程实录清华大学出版社.2011

[13]张世明等.C#编程语言基础和应用中国铁道出版.2011

[13]内格尔.C#高级编程清华大学出版社.2008

[14]Griffiths,Ian.Programing C#5.0 2012

[15]Joseph.Albahari and Ben.Albahari.C# 5.0 in a Nutshell O'Reilly Media .2012

[16]Jon Titus. ECN Technical Editor：”The Eclipse of stand[J]. Journal of Zhongkai Agrotechnical College”,Vol.19,No.2, 2006.

[17]Markus Aleksy,Axel Korthaus,“Martin Schader.Use Java and the CORBA realization distribute type system”, Journal of Pingxiang College,No.4,2005.

[18]PaulMiladjenovic.StockInvestingForDummies,2ndEdition.PublishedbyileyPublishing,Inc.111RiverSt.Hoboken,NJ07030-5774www.w-iley.comCopyrigh t(C)2006byWileyPublis

附录

| 指　标 |
| --- |
| 疑似剽窃文字表述 |

1. 系统的架构设计，可以分为五个模块：爬虫模块、网页解析模块、数据处理、浏览模块、预警模块。

## 表格检测结果

原文表格1：表4-4 FOCUS_LIST表结构

| 名称 | 说明 | 数据类型 | 长度 | 主键 | 外来键 |
| --- | --- | --- | --- | --- | --- |
| CODE | 股票代码 | VARCHAR2(20) | 20 | TRUE | TRUE |

相似表格1：表18tc "发布流程_用户组表\: OS_GROUP" \l 1用户组表

相似度：41.67%

来源：段丽娟_XX电网应急指挥管理系统设计与实施-段丽娟-《》-2015-09-22

原文表格2：表4-11 WARNING_QULAFICATION表结构

| 名称 | 说明 | 数据类型 | 长度 | 主键 | 外来键 |
| --- | --- | --- | --- | --- | --- |
| ID | 编号 | NUMBER | | TRUE | FALSE |
| SHOW_NAME | 显示名称 | VARCHAR2(200) | 200 | FALSE | FALSE |
| QUALIFICATION_CODE | 条件代码 | VARCHAR2(200) | 200 | FALSE | FALSE |
| TARGET_NUM | 目标数字 | NUMBER | | FALSE | FALSE |
| SIGN | 符号 | VARCHAR2(2) | 2 | FALSE | FALSE |

相似表格1：表4-12 供货商表

相似度：44.12%

来源：刘波201192120334成都多达商贸有限公司销售送货管理信息系统设计与实现-2修-刘波-《》-2013-09-01

相似表格2：表4-12 供货商表

相似度：44.12%

来源：刘波201192120334成都多达商贸有限公司销售送货管理信息系统设计与实现-3修-刘波-《》-2013-09-05

相似表格3：表4-12 供货商表

相似度：44.12%

来源：刘波-201192120334-成都多达商贸有限公司销售送货管理信息系统设计与实现-刘波-《》-2013-10-08

说明：1.指标是由系统根据《学术论文不端行为的界定标准》自动生成的。

2.红色文字表示文字复制部分; 黄色文字表示引用部分。

3.本报告单仅对您所选择比对资源范围内检测结果负责。

4.Email：amlc@cnki.net　http://e.weibo.com/u/3194559873　http://t.qq.com/CNKI_kycx