DOI:10.19850/j.cnki.2096-4706.2020.22.005

基于 HTTP 报文网络抓包软件的分析与设计

谢剑

（湖南信息职业技术学院，湖南 长沙 410200）

摘　要：网络信息获取在企业的决策制定中占据了重要的位置，结合“网络爬虫设计”课程内容，针对信息的自动抓取设计实现了一套基于 HTTP 报文交互的网络抓包软件。本软件设计并实现了多线程网络信息请求及解析、数据存储、参数配置、日志记录等功能，用于实时自动地获取特定网络信息并进行解析提取以及存储，以便后续统计分析。该软件抓住企业需求，具有较好的实用性。

关键词：网络信息；抓取解析；自动实时；实用性；抓包软件

中图分类号：TP391；TP311.5　　**文献标识码**：A　　**文章编号**：2096-4706（2020）22-0020-03

Analysis and Design of Network Packet Capture Software Based on HTTP Message

XIE Jian

（Hunan College of Information，Changsha 410200，China）

Abstract：Network information acquisition occupies an important position in the decision-making of enterprises. Combined with the content of the “Web Crawler Design” course，a set of network packet capture software based on HTTP message interaction is implemented for the design of automatic information capture. The software design and implementation of multithreading network information request and parsing，data storage，parameter configuration，log recording and other functions，which are used to automatically obtain specific network information in real time，extract and store it for subsequent statistical analysis. The software has good practicability by grasping the needs of enterprises.

Keywords：network information；grab and parsing；automatic and real-time；practicality；packet capture software

0 引　言

企业经营决策的制定离不开有效信息的支撑，特别是互联网时代的网络信息，信息量大、更新速度快，若仍采用人工方式收集，不仅需要花费较高的人力成本，而且还不能保证信息能够实时收集、实时上报，从而影响到企业运营策略的及时修正，进而影响到企业效益，这种情形在以赚取中间差价为盈利点的行业企业中尤为突出。软件通过程序方式自动抓取、解析指定网站数据并保存到数据库，从而取代传统的人工收集保存方式，在降低成本、提高效率的同时，也保证了数据的时效性。

本软件是笔者针对学校的专业课程“网络爬虫设计”而设计的一个案例，通过此案例的分析与设计，使学生加深对相关知识点的理解与应用。

1 系统需求分析与设计

1.1 系统需求分析

此系统的使用对象是企业员工。主要涉及的模块有网络请求模块、多线程处理模块、数据库访问模块、配置模块以及日志模块。网络请求模块用于向指定的网站发送网络请求，并对返回的数据进行解析，提取出需要的数据，为提高效率应采用多线程处理，可同时发起多个请求。数据库访问模块采用数据库连接池的方式实现，将解析出的数据入库保存，以便后续进行统计分析，指导决策。配置模块能够对连接的数据库信息、需要请求的网站、日志的级别、线程的数量等进行配置，程序根据配置的信息进行抓包及保存处理。日志模块主要用于记录网络交互过程、数据库操作等关键环节的日志信息，以便出现问题时能及时查找并解决。

1.2 开发环境及关键技术

该软件是基于 Widows 10 系统，采用 C++ 语言来实现的一个抓包软件。

数据存储采用 MySQL 数据库，MySQL 数据库是当前最流行的关系型数据库之一。与其他数据库将数据存放在一个大仓库内的存储方式不同，MySQL 数据库将数据存储在不同的表结构中，以此加快访问速度，提高数据访问的灵活性。同时 MySQL 数据库也是支持处理千万级数据记录的大型数据库，而且是开源的软件，不需要收取版权费用，因此本软件采用 MySQL 数据库来存储和管理数据。

软件的开发语言选取的是 C++，开发工具采用 Visual Studio，Visual Studio 是目前流行的 Windows 平台应用程序的集成开发环境，支持多种语言开发，使用起来方便快捷，可以有效提高开发效率。

软件关键技术在于多线程 HTTP 报文请求的异步处理及数据保存，因此采用线程池、数据库连接池、IO 完成端口相结合的技术手段实现。线程池是指程序根据配置文件的配置，在启动时就创建相对应的任务线程数，并保存到一个队列结构中，当有 HTTP 请求交互时，从线程中取出一个任务线程进行任务处理，完成后再放进队列中，这样可避免频繁的线程创建及释放，提高性能。数据库连接池同样也是程序根据配置在启

收稿日期：2020-10-27

动时就创建相应的数据库连接对象，并将这些连接对象保存到队列中，当需要进行数据保存时，从队列中取出一个连接对象进行数据库操作，完成后再放回队列中，避免频繁的创建及释放数据库连接对象，消耗资源。IO 完成端口是一种网络编程模型，是 Windows 平台下最高效的处理网络异步请求的方式，本软件使用IO完成端口来实现网络请求模块中异步通信功能。

1.3　系统详细设计

1.3.1　系统功能结构设计

软件主要由 5 个功能模块组成，功能结构图如图 1 所示，以下将从实现类图的角度对各个模块进行阐述说明。

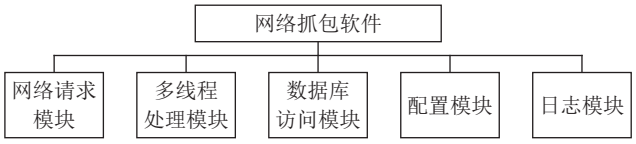
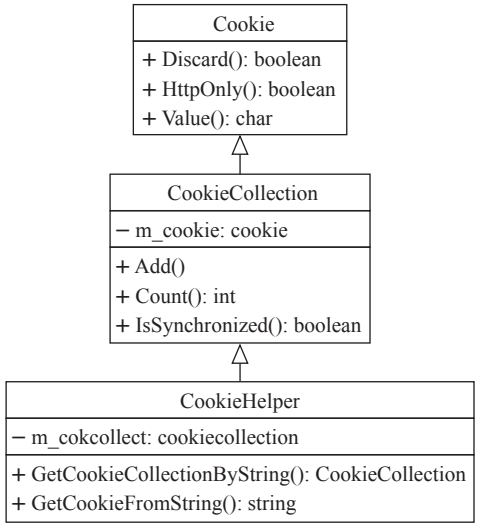
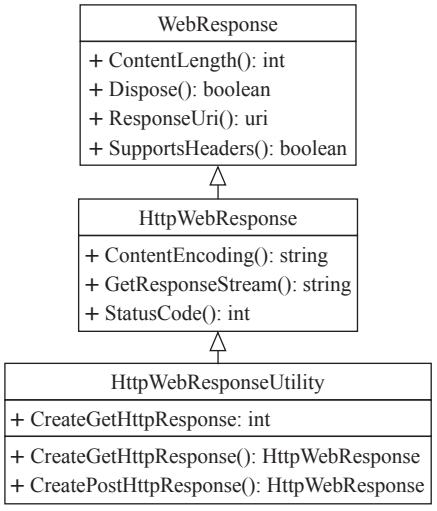


图 1　功能结构图

（1）网络请求模块。该模块主要用于进行 HTTP 请求的发送、接收以及数据的解析，其主要实现类图如图 2 所示。



（a）Cookie 相关类图



（b）请求相关类图

图 2　网络请求实现类图

Cookie 类是用于保存网络会话过程中服务器端生成的用来确定用户身份及会话连续性的信息的类，一个 Cookie 对象只保存一个键值对数据。CookieCollection 类是依赖 Cookie 类，用来保存并解析多个 Cookie 信息，并且对保存的数据可以进行增加或更新等操作，可随时计算出当前获取的 Cookie 数目并进行相关判断。CookieHelper 类依赖了上述两个类，负责从网络通信数据中截取出 Cookie 信息，生成对应的 Cookie 对象，并将新生成的对象放入到 CookieCollection 对象中，便于同一会话过程中的后续请求报文使用，保证服务器端的身份验证可以通过。

WebResponse 类是用于进行网络报文交互的基类，该类封装了报文交互相关的标准操作，可进行报文的发送和接收，并可获取到接收报文中的长度、报文头、报文体等相关内容。HttpWebResponse 类继承自基类 WebResponse，主要进行超文本传输协议（Hypertext transfer protocol，HTTP）报文交互，是互联网上一种主流的协议方式。该类提供一种无状态的交互方式，可实现客户端与服务器的数据、Office 文档、图片、音频、视频等文件的交互功能。HttpWebResponseUtinity 类则提供了 HTTP 协议定义的两种请求方式：Post 请求和 Get 请求。两种请求在参数传递和报文数据上会有一定差异，不同服务器有不同要求。通过对抓取报文的分析，可用该类模拟发送不同请求方式的报文并对返回数据的提取和解析工作，提取出所需要的信息进行保存，不同网站具体实现不同。

（2）多线程处理模块。该模块用于提供并发执行网络请求任务的能力，其实现类图如图 3 所示。

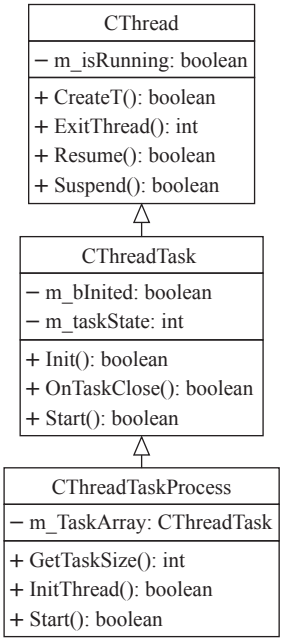


图 3　多线程处理模块实现类图

CThread 类是基础的线程类，该类用于在主程序中创建工作子线程，并可对正在执行的子线程执行暂停、恢复以及关闭操作。CThreadTask 类依赖 CThread 类，该类对基础类进行了封装，用来记录或获取子线程的工作状态（如运行、休眠等），并且与具体的网络请求任务相关联，当接收到任务请求时，就进行子线程的创建及相关初始化工作，并根据

指令对线程的工作状态进行修改。CThreadTaskProcess 类用来管理 CThreadTask 类，该类中实现了对线程数目的动态管理，根据实际需要进行线程数目的增减，实现一个线程池的功能，有利于资源的利用。同时该类还实现了同步锁，防止多个子线程在访问同一资源时出现死锁现象，提高了程序的稳定性，避免程序在出现此类问题时出现异常卡死的现象。

（3）数据库访问模块。该模块用于连接数据库并进行数据库相关操作，支持访问 MySQL 数据库，其实现类图如图 4 所示。

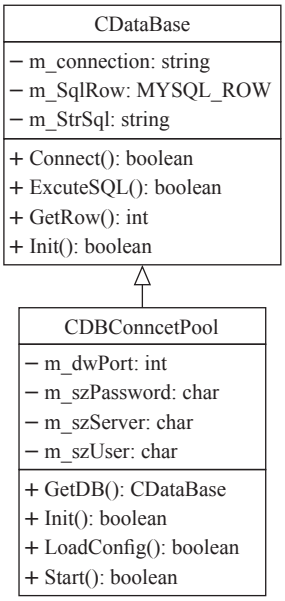


图 4 数据库访问模块类图

CDataBase 类是数据库的操作类，封装了相关的系统 API 函数，直接与数据库交互，负责相关初始化操作，并与具体的数据库进行连接，执行 SQL 语句，获得执行结果。CDBConncetPool 类是数据库连接池类，负责管理数据库连接对象队列，该类根据配置文件中的数据库相关配置（如：数据库地址、数据库名、数据库端口、数据库用户名及密码等）在程序启动时就创建好多个数据库连接对象，在程序退出时才释放这些连接资源。多线程情况下，每个线程只需要从该池中取出一个空闲的连接对象，当使用完时再把该连接对象放回队列中，供其他线程使用。

（4）配置及日志模块。配置模块支持数据库、访问网站、日志级别、初始线程数等多种数据的文件配置，配置项格式为键值对的形式，键值改动后需要重启程序才可生效。其实现类图如图 5 所示。

日志模块用于记录网络交互及数据库操的日志记录，以便出现问题时进行问题的分析查找。日志类以天为记录单位，跨天自动生成新文件，输出的每行日志记录时间精确到毫秒，同时会在每行头部显示该日志的级别，当配置为某一级别时，该级别及该级别以上日志会输出，该级别以下的日志不会输出。其实现类图如图 6 所示。

1.3.2 系统数据库设计

软件主要功能是网络信息的爬取及保存，不涉及其他业务需求，因此设计的数据库表只有信息保存表，用于存储抓取数据的相关信息，比如时间、信息类型、具体信息内容等。

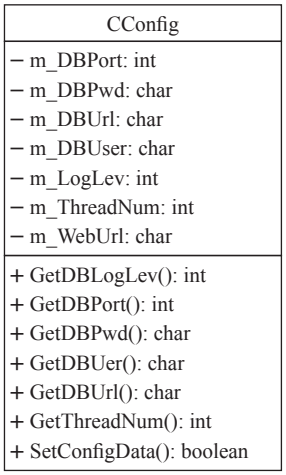


图 5 配置模块类图

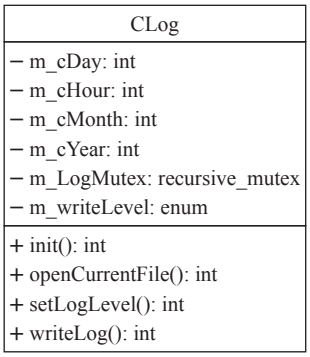


图 6 日志模块类图

2 结 论

本文详细描述了网络抓包软件的需求，开发环境及技术的选取、功能的详细设计与数据库表的设计，并已根据设计完成实现。若实际投入使用，企业可以通过此软件实现网络信息的自动收集及保存，可提高企业员工的工作效率，节约企业的人力成本。案例应用在了“网络爬虫设计”的课程中，通过讲解网络抓包软件的分析和设计过程，加深学生对知识点的理解及应用，同时提高学生系统分析与设计的能力。

参考文献：

[1] 吴洁明，方英兰 . 软件工程实例教程 [M]. 北京：清华大学出版社，2010.

[2] 王英英 .MySQL 8 从入门到精通 [M]. 北京：清华大学出版社，2019.

[3] 彭云鹃 . 应用设计模式的校园停车位收费系统的设计与实现 [J]. 冶金管理，2019（23）：79-81.

[4] 王佳珣 . 高校实验室知识管理系统用户需求分析与系统设计 [D]. 上海：华东理工大学，2013.

[5] 康昕宇，耿恒山，翟丹娜，等 . 基于 Android 的物流与财务管理系统的设计与实现 [J]. 计算机应用与软件，2016，33（8）：315-318.

作者简介：谢剑（1987.06—），男，汉族，湖南长沙人，就职于软件学院，教师，初级职称，硕士，研究方向：计算机应用、图像处理。