基于Scrapy的游艇主尺度数据库搭建与应用

赵立鑫[1]　孙承猛[2]　刘洋冰[3]
山东交通学院船舶与港口工程学院　山东威海　264200

摘　要：在游艇设计领域，合适母型的选择对设计效率提升至关重要。本文利用爬虫技术收集船舶主尺度及艇型参数，构建游艇初步设计阶段的主尺度数据库，并对数据展开初步分析，为后续主尺度预报工作奠定基础。采用Scrapy框架编写可拓展爬虫模块，抓取精艇网上在售游艇数据，详细阐述爬虫工作流程、设计思路及数据存储结构。基于搭建完成的数据库，按艇型和船体分类绘制游艇主尺度关系图，分析各参数间关系，为游艇初步设计提供参考依据。

关键词：游艇设计；Scrapy爬虫；主尺度预报

引言

传统游艇设计遵循螺旋上升的过程，设计师依据客户对游艇性能的需求不断迭代设计方案[1]。游艇初步设计的关键在于确定主尺度、重要艇型参数和型线数据，而选择合适母型是设计捷径[2]。现代造船业与计算机技术紧密相连，船舶数据库有助于设计人员查询检索历史数据，进而创新设计[3]。例如，Radojcic D V等[4]对美国海军Series 62船形数据构建高速水面舰艇数据库并优化，为舰船设计提供指导。与保密性高的型线数据不同，游艇主尺度数据相对较易获取，部分游艇交易平台网站会标注相关参数，这些数据可供设计参考。精艇网是其中的一个代表，网页中展示有各种类型的游艇相关数据。本文旨在研究如何通过爬虫技术对精艇网的游艇信息进行自动爬取搜集，构建主尺度数据库雏形，进一步处理筛选可用于分析游艇主尺度等参数间关系，为游艇初步设计提供助力。

一、爬虫技术

网络爬虫，从本质上来说，是一种依据系统化、自动化的运行模式，在广袤的网络世界中进行全面遍历，进而实现下载网络文档的程序或软件类型[5]。在当今数字化时代的各类应用程序场景中，它已然成为一种极为有效的手段，能够精准地针对特定网站的数据进行定向收集。其具体工作流程呈现出高度的逻辑性与规范性。

首先，爬虫会借助HTTP或HTTPS这两种广泛应用于网络通信的协议，向目标网站发起访问请求。当网站服务器接收到该请求后，会根据请求的合法性及相关规则，返回包含丰富网页内容的HTTP响应。此时，爬虫会运用其内置的解析模块对返回的HTML文档进行深度解析，从中精准提取出有价值的数据，并按照预先设定的规则将这些数据进行妥善存储。同时，爬虫还具备处理页面重定向的能力，当遇到网页页面发生重定向的情况时，它能够依据相应的机制进行正确的处理，确保数据获取的完整性和准确性。此外，在正常的运行过程中，爬虫通常会严格遵循目标网站所设定的robots.txt协议，该协议明确规定了网站中哪些部分允许被爬虫访问，哪些部分则禁止访问，这有助于维护网站的正常运行秩序以及保护网站的数据安全。

在Python编程语言的生态体系中，主流的爬虫架构主要包括Requests、Scrapy和Selenium等。杨健和陈伟[6]在其研究中对这三种爬虫技术进行了全面且深入的优劣对比分析。在充分考虑到后续具有拓展爬虫范围的实际需求这一关键因素后，本文最终决定选用Scrapy框架。Scrapy框架之所以脱颖而出，主要归因于其卓越的并发性能以及强大的拓展性。其并发性能使得在抓取精艇网上丰富多样的在售游艇数据时，能够高效地同时处理多个请求，极大地提高了数据获取的速度。而强大的拓展性则为后续进一步收集其他网站的游艇信息预留了充足的空间，确保了整个爬虫系统能够根据实际需求进行灵活的扩展和优化。本文利用Scrapy框架的主要任务在于精准地提取出对游艇设计具有关键意义的主尺度参数，以及其他众多有助于提升游艇设计质量的技术参数，同时还包括游艇内外布置的图像信息等各类数据，进而构

建起一个完整且具有高度实用价值的数据库，为后续的游艇设计工作提供全方位的数据支持。

二、基于Scrapy框架的爬虫设计

（一）Scrapy框架

Scrapy是开源网络爬虫框架，用于从结构化网站提取数据，其核心组件包括引擎、调度器、下载器、爬虫和项目管道等，基于Twisted异步网络库实现高性能并发。Scrapy提供Scrapy Shell，便于用户理解和调试爬取过程，广泛应用于各类爬取任务，灵活性强且社区活跃，是受欢迎的Python爬虫框架之一。

（二）Spider设计

本文利用Scrapy特性设计爬虫，Spider负责识别页面结构，将爬取数据抽象为Item，通过Pipline存储到数据库，Spider的爬虫逻辑如图1所示。

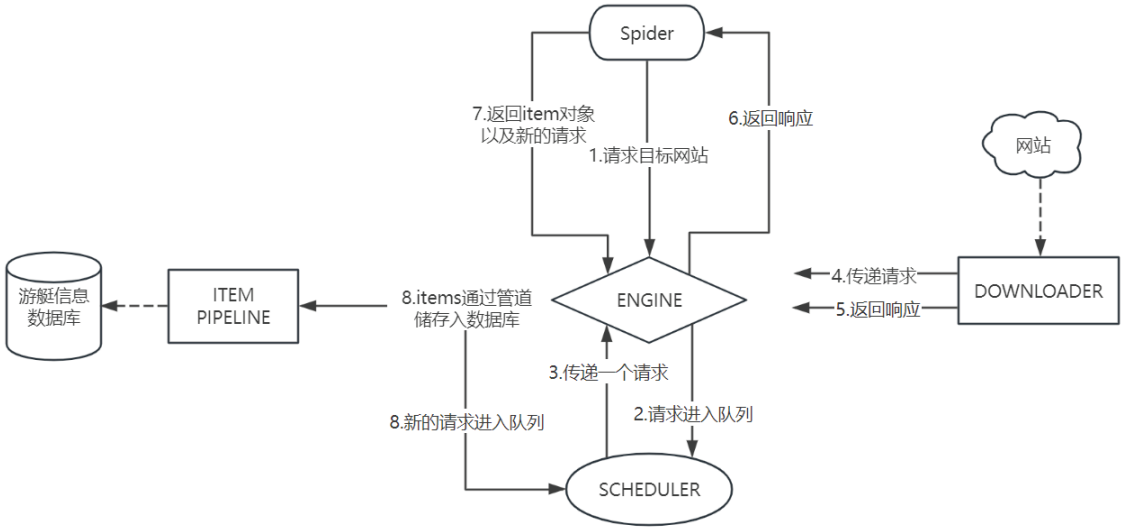

图1　Scrapy框架下的主尺度数据爬虫工作流程

Spider通过请求对象（request）访问网站来抓取网站内容，request对象到达下载器，下载器执行请求并返回相应对象（response）。response对象包含HTML页面的结构化内容。在Spider中，Scrapy使用选择器来提取HTML中的结构化数据，随后将结构化的内容封装成Item。Item是Scrapy的一个数据对象，它可以通过管道存储到数据库或者本地文件当中。管道通过Scrapy的引擎调度，用来对获取的数据进行持久化。

Scrapy选择器机制的核心是Xpath和CSS的表达式。XML Path Language（简称Xpath）种用在XML文档中选择节点的语言，也可以与HTML一起使用。层叠样式表（简称CSS）一种用来为结构化文档添加字体、大小、颜色以及间距的计算机语言。Xpath选择器和CSS选择器各有优缺点，在具体使用时，可以将两者配合使用。

（三）Item以及Pipeline设计

在游艇设计中，常用的主尺度参数有艇长、艇宽、吃水和排水量等，游艇主尺度数据对象依据游艇型号区分；游艇的总体布局也是游艇设计时非常重要的内容之一，在进行爬虫时，这些数据也是有价值的。

游艇的内外布置往往以图像的形式呈现，爬取这一部分数据需要设计专门用于存储图片的数据对象。爬虫并不会直接爬取图片，而是将图片的存储路径，图片的URL（Uniform Resource Locator，统一资源定位符的简称）进行整合。在Spider将数据传递给管道（Pipeline）之后再进行下载。因而，从Spider中传递过来的数据，需要对图像和主尺度数据进行两种不同的管道设计。

三、游艇主尺度数据库

（一）数据爬取结果

爬虫将提取的游艇主尺度信息以及乘员数等在游艇初步设计阶段对游艇主尺度、型线以及总舱室布局有影响的数据存储到CSV文件当中。在游艇数据总量并不大的情况下，直接使用CSV文件存储数据。Python的数据科学库Pandas库可以直接读取CSV文件进行数据分析。游艇的图像存储在本地文件夹中和游艇名称一一对应。

通过爬虫从精艇网爬取了200余艘游艇信息。按照游艇长度、游艇功能以及片体数量等划分方式，给出了饼状图分布，如图2所示。

（二）数据结果应用

对数据库中的主尺度信息，进行简单的分析。如图3（d）所示，游艇长宽的关系呈一条斜率一定的直线，所以通过长宽比观察艇重和吃水的分布如图3（b）、3（c）所示。吃水随长宽比的分布较为分散，而艇重随长宽比的分布更贴近三次多项式。艇重和吃水的关系如3（a）所示，也比较分散。

结束语

本文通过Scrapy爬虫框架实现精艇网在售游艇数据

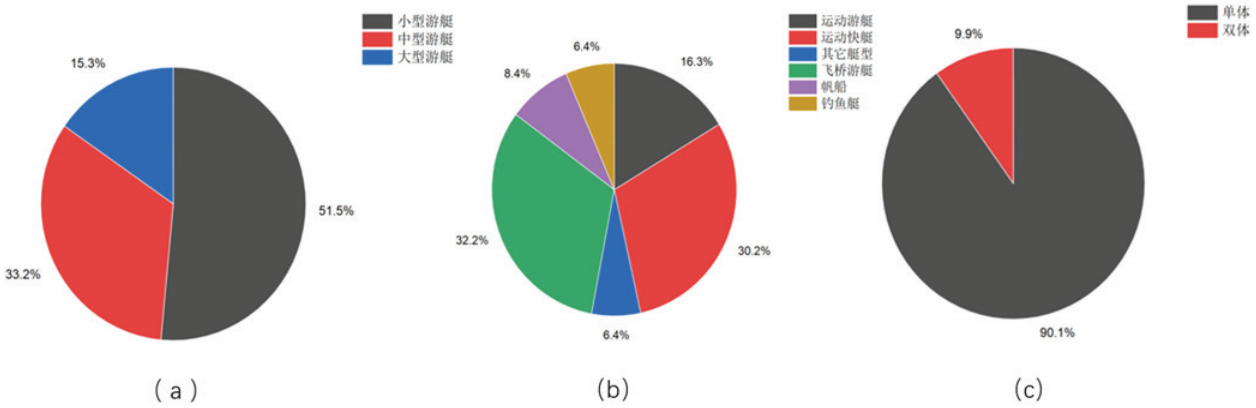

图2　游艇数据分类

（a）按游艇艇长分类；（b）按游艇功能分类；（c）按片体数量分类

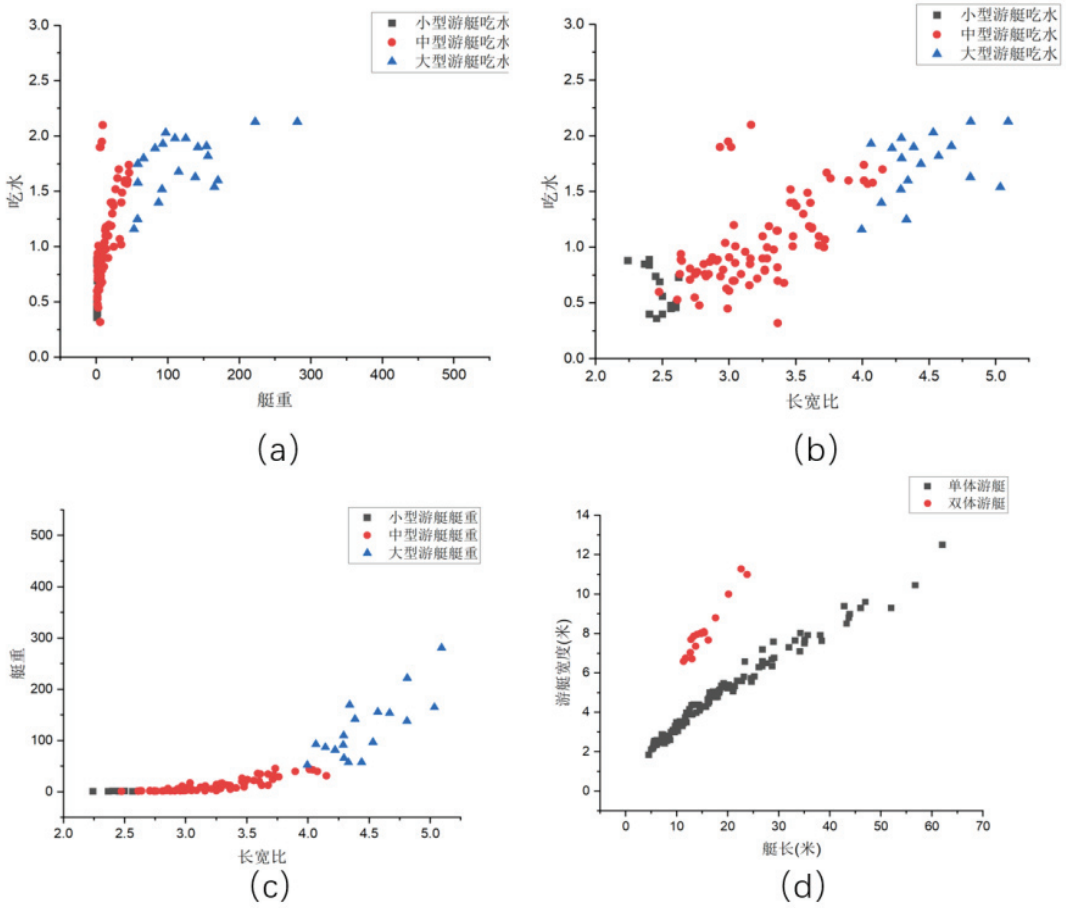

图3　游艇主尺度关系图

（a）艇重吃水散点图；（b）长宽比吃水散点图；（c）长宽比艇重散点图；（d）艇长艇宽散点图

的爬取，从中提取游艇主尺度数据组成游艇主尺度数据库，并对游艇主尺度信息进行了初步的分析。

但是数据库中的信息还不能够完全支持游艇初步设计。单就主尺度信息而言，其中的数学关系有待进一步分析。为实现辅助游艇初步设计的目标，可以数据库为基础进行回归分析或者建立神经网络模型拟合隐藏在数据库中的非线性关系。

主尺度数据库可以作为型线设计的基础，通过参数化建模或者数学船型等方法将主尺度数据库拓展呈游艇型线数据库，这将是后续研究方向。

参考文献

[1] 薄林.现代游艇设计和应用[M].哈尔滨：哈尔滨工程大学出版社，2016.

[2] 周志成.母型船设计模式在海军舰艇数据库构建中的应用[J].舰船科学技术，2016，38（02）：118-120.

[3] 严珂.专家系统和数据库技术在船舶设计的应用[J].舰船科学技术，2021，43（18）：4-6.

[4]Radojcic D V, Zgradic A B, Kalajdzic M D, et al. Resistance and Trim Modeling of a Systematic Planing Hull Series 62 (with 12.5° , 25° , and 30° Deadrise Angles) Using Artificial Neural Networks, Part 1: The Database[J]. Journal of Ship Production and Design, The Society of Naval Architects and Marine Engineers, 2017, 33(03): 179-191.

[5]Singh Ahuja M, Singh Bal D J, et al. Web Crawler: Extracting the Web Data[J]. International Journal of Computer Trends and Technology, Seventh Sense Research Group Journals, 2014, 13(3): 132-137.

[6] 杨健，陈伟.基于Python的三种网络爬虫技术研究[J].软件工程，2023，26（02）：24-27，19.