

题 目： 网络爬虫的设计与实现

网络爬虫的设计与实现

摘 要 随着近二十年来编程语言的高速发展，使得之前没有那么热门的一种高级语言 Python 又活跃了起来。其实 Python 是一种比 Java 还要古老的面向对象的动态类型的编程语言，2010 年之前关注度不是特别的高，但最近几年在大数据和机器学习方面得到了广泛的应用，相对于 C、C++、Java 语言等，它具有许多独到的优势，因为其简单的语法结构使得它易于学习和阅读，也易于维护，广泛的标准库使得它更加的方便灵活。所以由于种种原因，使它能够更好地去编写爬虫代码。时代不断的进步使得爬虫与反爬虫行业得到了前所未有的发展，在方便了整合数据的同时还让人们对这个行业有了新的认识。网络爬虫不仅能像搜索引擎一样得到信息，对目标网页 HTML 进行爬取。还能够爬取一些特定的你所需要的东西。本项目实现了基于 Python 的爬虫程序，针对 58 同城网站进行爬取。并在系统实现过程中解决了数据分析和页面可视化等一系列问题。

关键词 Python；网络爬虫；HTML；58 同城；

Design and Implementation of Web Crawler

Abstract With the rapid development of programming languages in the past two decades, Python, a high-level language not so popular before, has become active again. Python is actually a kind of more ancient than Java object-oriented dynamic type of programming language, attention before 2010 are not particularly high, but in recent years in the big data has been widely used and machine learning aspect, relative to C, C++, Java language, etc., it has many unique advantages, because of its simple syntax structure makes it easy to learn and read, and easy to maintain, and a wide range of standard library makes it more convenient and flexible. So for a variety of reasons, so that it can be better to write crawler code. With the continuous progress of The Times, the crawler and anti-crawler industries have achieved unprecedented development, which not only facilitates the integration of data, but also enables people to have a new understanding of this industry. Web crawler can not only get information like a search engine, but also crawl the HTML of the target Web page. You can also crawl for certain things you need. This project implements a crawler based on Python, which crawls 58 city Websites. In the process of system implementation, a series of problems such as data analysis and page visualization are solved.

Keywords Python; Web crawler; HTML; 58city

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：

<https://d.book118.com/548127123074006132>