基于 Python 的数据抓取和用户特征分析

甘长笑，孙立太

武汉东湖学院　湖北武汉

【摘要】在最近几年，在大数据的快速发展中，数据成为了网络产业最宝贵的财富。特别是在如今 B2C 的时代，资料的使用越来越有意义。大量的资料是有很高的科研价值的。在网络产业中，数据挖掘有着举足轻重的作用。由于社会网路的日益流行，社会网路的传播与传播的速率日益提高。有许多资料是由直接或间接提供的。在任何一个纵向的研究中，都需要对目标网站进行实时捕捉和分析，并将这些信息传递到目标使用者。文章论述了 Web Crawler（Web Crawler）的基本理论和构成，并对其构成进行了详细的描述，并将淘宝作为一个有代表性的 app 进行了数据收集和分析，从而对淘宝的各种特性进行了优化。

【关键词】Python 爬虫；数据分析；用户特征分析

Data Capture and Analysis Based on Python

Changxiao Gan, Litai Sun

Wuhan Donghu University, Wuhan, Hubei

【**Abstract**】In recent years, in the rapid development of big data, data has become the most valuable wealth of the network industry. Especially in today's B2C era, the use of data is becoming more and more meaningful. A large amount of data is of high scientific value. Data mining plays an important role in network industry. Due to the increasing popularity of social networks, the spread and speed of social networks are increasing day by day. Much information is provided directly or indirectly. In any longitudinal study, it is necessary to capture and analyze the target website in real time and transmit the information to the target users. This paper discusses the basic theory and constitution of Web Crawler (Web Crawler), and gives a detailed description of its constitution, and takes Taobao as a representative app for data collection and analysis, so as to optimize the various characteristics of Taobao.

【**Keywords**】Python; Data analysis; User Feature Analysis

1 绪论

1.1 课题背景、意义和目标

（1）课题背景

大数据时代的来临，信息技术的飞速发展，特别是物理信息系统、互联网、云计算、社交等领域的飞速发展，使得大数据无所不在，已经成为信息社会的一项宝贵资源。但这也是一个很大的问题。数据的可用性是一个很大的问题。不管是在商业或财务方面，都有越来越多的数据依赖性。伴随着大量的资料，也出现了一些劣质的资料。由于资料的获取问题，给资讯社会造成极大的危害，已经成为业界和业界所关心的问题。

所以，在网络中，数据挖掘是非常有必要的。在网上淘宝等网上商城越来越流行，网购也逐渐成为了人们的必需品。但是，对于各个群体的属性、用户特征以及个体的喜好，我们却知之甚少。在淘宝和部分商户中，挖掘用户信息，分析用户属性特性是很有必要的。人们对网络大数据的巨大价值进行了深度挖掘和充分利用，带来了巨大的机遇。我们可以比较淘宝内的用户各项数据，分析用户特征和属性，商家根据结果做出更好的活动。用户们就可以用手机查看到更偏向自己喜好的商品，也可从商品分类中查看自己的特色爱好，为人们的生活带来更多便利和乐趣，帮助人们轻松购物，并快速了解用户的特点。

（2）课题的意义

淘宝软件凭借其实时性和快速性的优势，逐渐受到公众的关注。特别是近年来，各种购物 app 的

出现促进了主要购物平台的发展。淘宝等商品在网络上也是必不可少的。网络商城的兴起是为适应网络消费而产生的。用户只需拥有自己的账户，就可以随时添加购买自己喜欢的商品。

同时，这种网上购物的普及也带来了一些弊端。初级阶段商家盲目推送一些劣质广告来迷惑用户，在不了解用户的情况下追随潮流，在浏览软件时用户无法得到关键信息。同时，还能实时查询到淘宝上的最新热点产品，让你更好地掌握市场动向，并对其进行追踪、分析，掌握其喜好特征及特征，便于查询。所以，资料的处理能够有效地推动人们的日常生活，使人们的日常工作更加便捷和快速。

2 开发工具和相关技术介绍

2.1 开发工具简介

此 APP 以 Eclipse 为平台，以 Java 和 Python 为基础，以 Django 架构构建了系统，利用 Scrapy 进行数据采集、对关键字进行预处理。

2.2 Python 简介

Python 是一门具有简洁的文法和丰富的类库的解释型电脑程序设计语言。Python 是一个注重开发的快速和清晰的编程语言。从一个简单的指令指令到一个高级的面向目标的应用，都可以使用这个软件。Python 还被认为是一种很好的编程工具，因为 Python 是免费的，面向对象的，可扩展的，而且有一个很好的代码规范[2]。当需要更高的效能时，可以使用 C/C++覆盖并将其包装成一个可供 Python 使用的扩充类库。在网络设计和系统的管理工作中有着广阔的应用前景。

Python 为我们提供了一个很好的基础程式码，包含很多东西，例如网络，文件，图形用户界面，数据库和文本。这种物质叫做“含电池”。很多功能都可以在 Python 中直接使用，不一定要重新写。

python 中的许多第三方类类库，都是别人为你提供的，你可以直接利用它们。当然，如果你编写了一个经过良好包装的程序，那么你也可以将其作为第三方程序的使用。

2.3 Scrapy 爬虫简介

Scrapy 是一款快捷、先进的网页捕捉与网页爬行架构。该软件将访问站点并从网页中获得结构性的资料。从数据采集到监测，再到自动化检测，都可以使用。Scrapy 是用 Python 完全实现的，它是开放源代码。在 GitHub 中，可以使用 Linux，Windows，MAC 和 BSD 等多种应用环境，利用 twisted 的异步网络资源，在此基础上，用户可以方便地自定义和编写多个应用模块来完成网站的爬行。

Scrapy 利用 twisted 的非同步的网路资料库来进行网路通讯。

2.4 Django 框架

Django 是 Python 编写的开源 Web 程序架构。该系统的主要功能是 MVC 的建模方式，包括 MVC、View 和 C。Python 在 Django 中广泛应用，甚至还包含了概要和资料模式。

Django 处理流程：

（1）加载配置；

（2）启动；

（3）处理 Request；

（4）创建 Response；

（5）处理 Response。

总的来说，在 fastcgi 中，Django 的两项工作是：Request 和 Response，而他们的中心是“urls 分析”，“模板技术”以及“ORM 技术”，这些都是稍后分析的。

3 环境搭建及安装

基于 Python Cash 的构建和设置

（1）首先是到 www.python.org/download/下载 Python 的最新版，这是一个很好的例子。

（2）安装下一步的程序。

（3）将一个安装文件夹加入到一个环境变数中，例如在 pth System 中增加 Python 的安装文件夹。

（4）通过 cmd 的命令，输入 Python，然后点击下面的图片。

（5）Hello World！按照很多资料上写的，输入 print 'Hello World!'

（6）Python 的安装完成了，不过该如何进行，用这样的指令来完成？还有 IDE 也可以使用，可以通过安装在 Eclipse 上的插件集合来安装 Aptana Studio，它可以为 Python 提供支持。请访问 http://aptana.com/products/studio3/download，下载并运行也可以在你自己装上 Eclipse 之后自行寻找 PyDev 的插件。

（7）在你的设计中，先对这个 IDE 进行一些环境的设置。

（8）新建一个项目

这个工程已经完成，Python 创建了一个新的 Python，然后键入了这个程序，然后运行并展示了它的效果。就这么简单。

4 基于 Python 的媒体数据抓取

4.1 数据抓取（网络爬虫）技术

在分析使用者的特征和信息推荐等方面，本文提出了基于多媒体信息的信息收集与处理技术。本章主要从数据收集与处理的角度出发，探讨了基于网络环境下的数据收集与处理中的Web数据分析方法，并在此基础上进一步完善了数据的可用性和准确性[3]。

Web 爬行器本质上是一种以 Web 为基础的软件。该算法从最初的一组页面出发，对 Web 数据进行了搜索和自动采集。在网页被开启时，会对 HTML 标签的构造进行解析，得到相关的资讯，并根据所设定的搜寻战略，选取下一个网页。从原理上讲，可以覆盖整个互联网[3]，并为整个网络指定合适的初始搜索策略。它的性能极大地影响了搜索引擎网站的规模[2]。

在研究者们的不断探索和实验中，目前的爬虫技术已由最初的单一的以全网为基础的爬行方式发展成为一种可以适应多种需要的多收集技术。

文章以某一具体的站点为研究对象，以搜集用户所需的资料为研究对象。相对于以前的爬行器，它的搜集对象更为清晰，而且可以为使用者提供所需要的资讯。这样既节约了大量的硬件和网路资源，又减少了网页数目，使网页得以快速地更新。同时，由于能够很好的适应特殊群体对特殊领域的需要，也是目前的一个重要课题。

4.2 媒体数据抓取

当 Python 抓获的构建完成之后，您就必须先使用 Python 的 Pip 软件包管理软件 pip，或 easy_install 来安装 Scrapy 抓取。为特定的网站提供一个获取该网站的内容。

Spider 是该计划的中心内容。在此类别中，我们将会定义一个爬虫对象（域名，URL）和爬行法则。在 scripy 的正式文件中，指南是以 basepiper 为基础的，但是 basepiper 仅能够获取一个 URL 的清单，而无法按照最初的 URL 来进行。但是，在 basespider 以外，很多可以直接从 spider 中继承，例如 scratch contrib。“是蜘蛛。”爬虫

在 East 中，在 money/spider 目录中新建一个 East. Py，然后完成该代码。

（4）存储数据

当一个爬行器获得了数据之后，我们必须把它保存到一个数据库里。我们以前已经说过，这个动作必须通过工程管线来完成。常用的运算如下：

a、清理数据；

b、证书解析的数据（检查项目是否包含必要的字段）；

c、检查是否为重复数据（如果重复则删除）；

d、将解析后的数据存储在数据库中。

4.配置文件

在你执行爬行程序前，你还必须在设定中设定[6]，并且将某些组态资讯加入 py。

至此，该爬虫类的捕捉工作已结束。在指令列中运行“scratch crawl East uMoney”的“蜘蛛”，让它爬起来！

5 总结与展望

本研究以 Python 的数据采集与使用者特性为基础，利用 Python 的爬行器捕捉并进行分析。通过需求分析，模块划分，技术调试分析和设计，从而达到博文的目的。Bowenworld 能够帮助前端的使用者迅速地浏览和解析资料。后台管理人员可以方便地获取数据，保存数据，分析和管理数据。

博文天下让读者们更便捷和快速。利用博文世界，读者在不需要读到海量的资料的情况下，就能随时浏览到他们的日志，并理解他们的特征。对不同的社会网络进行采集和分析，得出了不同的用户特性，并得出了不同的社会网络环境下的用户特性。通过对网络平台的研究，可以了解到各大平台的用户的不同需要，以及对其进行股票的研究以及对其进行统计和统计。

博文天下的总体功能还是很基础的，要完善的部分还是非常多的。还需对编码进行最优。因为时间的关系，在很多地方都没有重要的信息，所以很多地方都要用到人工智能。没有实施使用者互动。对于博文的不足和不足，我会在以后的研究中不断地学习和改进。

在以后的工作中，努力使其具有更好的自动化捕捉能力、更好的检测准确率、更多的抓持数量和

速率，从而增强了整体的可操作性。

参考文献

[1] 张华平,高凯,黄河燕,赵燕平.大数据搜索与挖掘[M].科学出版社.2014.3-6

[2] 龙香好.基于网络爬虫技术的数据抓取程序的设计[J].技术与市场,2021,28(10):41-43.

[3] 徐志,金伟.Python 爬虫技术的网页数据抓取与分析[J].数字技术与应用,2020,38(10):30-32.

[4] 任洛漪.基于 Scrapy 的商务网站数据抓取[J].信息与电脑(理论版),2018(19):56-57.

收稿日期：2022 年 8 月 10 日

出刊日期：2022 年 9 月 25 日

引用本文：甘长笑，孙立太，基于 Python 的数据抓取和用户特征分析[J]. 工程学研究, 2022, 1(3) : 49-52

DOI: 10.12208/j.jer.20220062

检索信息：RCCSE 权威核心学术期刊数据库、中国知网（CNKI Scholar）、万方数据 （WANFANG DATA）、Google Scholar 等数据库收录期刊

版权声明：©2022 作者与开放获取期刊研究中心(OAJRC)所有。本文章按照知识共享署名许可条款发表。https://creativecommons.org/licenses/by/4.0/

OPEN ACCESS