Python 语言在 Web 数据挖掘中的应用

聂莉娟[1]　方志伟[1]　赵心宇[2]
（1. 金肯职业技术学院　江苏省南京市　210000　2. 视若飞信息科技（上海）有限公司　江苏省南京市　210000）

摘　要：本文重点阐述了 Python 语言特征及其在 Web 中的应用，指出了 Web 数据架构，提出了基于 Python 语言的 Web 数据挖掘与分析方法。

关键词：Python 语言；Web；数据挖掘

1 引言

在现代科学技术水平发展日新月异的背景下，大数据技术与云计算技术得到了快速发展，于此同时越来越多的行业领域在日常生产与运营过程中均会出现海量的数据信息，怎样有效的处理这些海量的数据信息，并从中获取拥有较高价值的信息，慢慢成为了现代人员探究的关键所在。随着这种诉求的越来越强烈，数据挖掘技术慢慢浮出水面，正式进入到人们视野当中。随着数据挖掘技术的不断发展，Python 语言在其中占据了极其关键的地位，并慢慢发展成为了运用非常广泛的数据挖掘技术。当下 Web 信息的呈现出指数级增长，要想实现对庞大数据信息的准确分析与筛查，同时充分发挥数据信息的价值，必须要加强数据挖掘技术的改革与创新。Python 是一个面向对象的开源程序设计语言，相比较于 C 语言以及 C++ 等其它编程语言，Python 语言的语法结构较为简单，并且因为 Python 语言是建立在 Guido van Rossum 基础之上开发的，这使得 Python 语言具备较多类型的库与 API[1]。正确巧妙运用 Python 语言中的不同工具库，比如说 sklearn 工具库等，可以逐步改善 Python 语言的运用价值，为强化 Web 数据挖掘速度与效果提供有力支持，以此来提升社会生产力 [2]。本文重点阐述了 Python 语言特征及其在 Web 中的应用，指出了 Web 数据架构，提出了基于 Python 语言的 Web 数据挖掘与分析方法，不断改善 Web 数据挖掘效率与效果。

2 Python语言特征

Python 语言是在现代社会环境下产生的一种新型程序编程语言，该语言以第四代程序为载体，面向的关键性目标是交互性与分析性对象，在 Web 技术以及非 Web 技术中有着广泛而深入的应用，并在其中发挥了至关重要的作用 [3]。比如， Google 公司以及豆瓣网等许多网站在进行程序编写过程中都使用了 Python 语言。Python 是现代非常重要的脚本类语言之一，有着比其它编程语言更好的代码开发率，结合第三方数据库，只需要依托于少量的代码便能够展现出其十分强悍的性能 [4]。

在 Python 语言当中只存在少量的可读性代码，相比之下，当 C 语言要达到与其一样的性能时，自身包含的可读性代码通常在 20% 左右 [5]。除此之外，Python 语言的实际使用优势还体现在其它方面，比如说编程过程中省略了括号。当存在 begin…end…类型语句时，在 Python 语言中科学合理运用冒号便能够完成代码的不同分层。例如在面向某个条件语句 ifTrue：print“Yes”中，句中 ifTrue 后面便运用了冒号，表示下一行 print 为下部基础语句，这种情况下要达到既定的条件才会往下执行。

3 Python语言在Web中的运用优势

相比较于其它编程语言，Python 语言拥有非常好的跨平台与开源性优势，将其应用到 Web 程序中以后，使其优势得到进一步扩大 [6]。Python 语言运用的 WSGI 模型涵盖在该语言服务器范畴之中，期间使用的各种程序和中间层都建立在官方标准之上，尽管如此也存在一定的不足，即无法准确辨别异步模型。除此之外，在 Python 语言快速发展的背景下，其运用优势愈加显著，慢慢转变成为智能手机游戏产业编程的关键性语言，逐步取代了传统的 C 语言与 C++ 等传统编程语言。

将 Python 语言和云计算技术充分融合到一起，进一步推动了基层程序从传统方式转型为虚拟化模式。充分发挥云计算技术运用优势与价值，为使用者提供更加高效、便捷的信息资源服务，具体涵盖了 Saas/ Iaas 等，例如，在线办公平台易度平台建设过程中便融入了较多的 Python 语言。需要特别留意的是，Python 语言可基于标准化数据库科学合理弥补大数据中存在的多个问题，实现对大数据的处理、转化以及分析等，之后归纳与总结出大数据的本质属性，充分展现出其结构方式的各项要求。

4 Web数据的主要设计架构

4.1 Django

● 基金项目：江苏省职教学会 2021-2022 年度职业教育研究课题《民办高职院校“专企融合、岗位分级实现梯队式教育”人才培养模式的实践研究》（XHYBLX2021010）。

Django 是一种应用非常普遍的 Python Web 开发结构。这种 Web 开发结构当中，自身表现出较高的开源性，包含有不同类型的组件，可以对存储、显示界面以及映射关系等性能开展动态管理。Django 架构的设计通常建立在 DRY 原则基础之上，由于本身包含有单独的、轻量性的 Web 服务器，只需要投入小部分时间便可以改善 Web 技术开发与使用效果 [7]。

在进行 Django 设计时，需要严格依照遵守 MVC 模式原则，共计包含三个方面，即控制、视图以及模型。从应用程度方面来分析，模型层处于最底层，该层的关键性任务便是科学合理的处理与数据紧密相连的事务，例如对多种类型数据的查验、对庞大数据的规范存储等。因为在 Django 中用户发出的命令均要依托于基本框架来处理与实现，所以该层也被叫做模块层。模块层的主要工作职责便是全面展现各类数据，同时需要实现对存取模块以及对模块的科学合理使用。设计人员在选择应用模块语言实现对 HTML 界面渲染时，需要赋予模块层相应的数据，采用个性化的模板得到需要的渲染效果。视图层是应用程序中十分重要的一部分，其关键性工作任务便是展现界面以及相关文档等信息之中的主要数据。

Django 的实际操作流程主要包含以下几个部分：

（1）浏览器将 HTTP 请求发送到 Web 服务器。

（2）当 HTTP 请求送达到 Web 服务器以后，服务器转变为运用 Django。

（3）Django 会下达指令要求中间层根据 URLconf 模型对各项数据进行匹配，同时与相关函数联系起来；函数以不同模板与模型为基础，依照具体需求产生不同的响应；之后中间层将产生的各种响应转化为 HTTP 响应，最后将该响应反馈给 Web 服务器。

（4）Web 服务器在接收到反馈响应以后，将其发送到使用者浏览器。

4.2 Pyramid

Pyramid 属于开源架构中的一种，在执行工作过程中表现出较高的效率，能够有效节省较多的设计时间，提高设计人员开发效率与效果。Pyramid 不仅涵盖了 Python、Perl 等特性，同时还拥有快捷高效的开发性能 [8]。

4.3 Flask

Flask 一般广泛使用在轻量级 Web 当中。通常情况下，轻量级 Web 服务器的网关接口主要是 Werkzeug，选择的模块引擎主要是 jinja2，并使用 BSD 对其进行授权 [9]。事实上，单一的 Flask 既不包含有抽象的数据库，也没有对各项表单进行查验的基层性能，在实际运行过程中，Flask 主要是依托于其它数据库来实现相关功能。Flask 架构表现出显著的拓展性特征，能够十分方便的增设相关功能，从而更好达到使用需求。

4.4 CherryPy

CherryPy 是以 Python 面向对象的 HTTP 架构为载体，主要服务的群体是 Python 设计人员。设计人员在开展 Web 技术开发与应用过程中，可以引入 CherryPy 技术，然而需要特别注意的是，CherryPy 自身不存在完善的语言体系。在 CherryPy 当中包含有 Web 服务器，在这种情况下，使用者可以直接使用 CherryPy 自带的服务器，免去了搭建 Web 服务器的麻烦，便能够依托于内置程序实现运行。服务器的主要职能体现在以下两个部分：首先，将从底层传输来的 TCP 套接字信息统一转变为 HTTP 请求，之后将转换好的信息发送到处理程序中。其次，当接受到从上层软件发送来的数据时，及时将这些数据变化为 HTTP 响应，同时将信息传输到基层的 TCP 套接字。

4.5 TurboGear

相比较于其它 Web 设计架构，TurboGear 不能够单独存在，必须要依附在相关架构之上，致力于充分挖掘与发挥各个架构的优势与价值。依托于 TurboGear 设计架构，开发者可以先从基础性的文件服务方面入手，之后不断拓展到全栈式服务 [10]。

4.6 Django和Pyramid、Flask之间的差异

Flask 应用的场景主要是一些较为简单的小应用，属于微框架范畴，相比之下，Diango 以及 Pyramid 主要应用在一些大规模功能当中。尽管 Diango 与 Pyramid 面向的对象较为相似，但是在灵活性以及拓展性方面两者又存在显著差异。Pyramid 具有较高的灵活性，开发人员在实际开展设计工作过程中能够灵活运用不同工具进行选取。例如基于 URL 架构以及数据库等选择流程。Diango 能够为 Web 开发者提供多种处理方法，并且由于其包含各式各样的模板，可以有效提高设计便捷性与效率。

在 Diango 设计架构中涵盖有 ORM 模块，开发人员在进行 Pyrami 以及 dFlask 设计时，可以依照实际需求灵活使用数据储存工具，在 ORM 模块当中 SQLAIchemy 模块拥有十分广泛的应用范畴，也可以选择 MongoDB 和 Dynamo 等其它模块。

Diango 设计架构能够为开发人员提供一系列完整的服务，在实际开展设计过程中，开发人员不再需要花费较多时间来确定基础的设备架构等事务。Diango 架构包含基础模板管理、表单、查验与数据库等深层次的创建性能。Pyramid 主要体现在查验与路由两个方面，基础模块管理以及数据库合理使用通常需要由第三方来实现，也可以基于 Pyramid 与 Flask 进行相关建设，开发人员必须要亲自完成构建选择，这样才能够更好发挥其灵活特性。

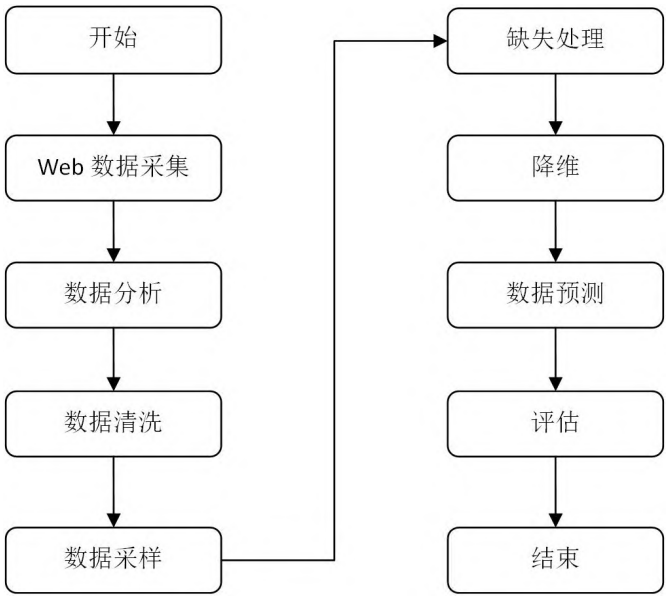

图 1：Web 数据挖掘的工作流程

5 Python语言背景下Web数据的挖掘与探究

当下在开展设计环节脚本过程中，Python 语言的应用十分广泛，由于 Python 语言表现出显著的交互性、解释性等特性，从而使得依托于 Python 语言编写完成的程序代码表现出较高的可读性，除了该优势之外，使用 Python 语言完成的结构代码还有助于降低后期程序二次开发与维护的难度，有效改善后续工作开展的效率与效果。除此之外，相比较于 C 语言、C++ 编程语言，Python 语言对初学者表现出更高的诚意，能够为其提供通俗易懂的运用环境，使其在实际编程过程中拥有更高的灵活性，正因为如此 Python 语言才得到了大范围运用，越来越多的开发人员习惯使用 Python 语言来完成编程任务。

5.1 Web数据挖掘概述

（1）Web 数据挖掘的定义。在许多学者看来，也可以将 Web 数据挖掘叫做 Web 信息挖掘、网络信息挖掘等，是指通过模拟用户正常的浏览器行为，并设置一定的规则，从而获取 Web 页面指定信息，实质上是一种联系数据挖掘与 Web 的新型技术 [11]。Web 数据挖掘技术的重要功能便是能够在较短时间内从庞大的杂乱无章数据信息中寻找出有用的信息，同时依托于数据转化、分析以及建模等一系列操作，对获得的数据信息开展深入分析与研究，紧接着依照分析结果对现状进行审视与考察，同时给出有效的预测性决策，由此可见 Web 数据挖掘表现出极高的科研价值与商业发展前景，如图 1 所示为 Web 数据挖掘的工作流程图。

（2）Web 数据挖掘的类别与特性。Web 数据挖掘开展的最终目标是将网页超链接、网络界面内容与 Web 应用日志等中的信息整合起来，并对其开展深入分析与挖掘，进而得到较多有用的信息，所以依照 Web 数据挖掘目的的差异和数据类型的不同，通常将 Web 数据挖掘分成以下三个部分，分别是 Web 结构挖掘、Web 应用挖掘与 Web 内容挖掘。上述三种挖掘方式中，Web 内容挖掘涵盖了文本挖掘和多媒体挖掘两种类别。因为 Web 本身拥有显著的特征，这使得 Web 数据挖掘也表现出较强的繁琐性、动态性与异构性等特征。在具体运用过程中，Web 数据挖掘涵盖了五大部分，分别是 Web 资源收集、Web 数据简单处理、Web 数据变幻和整合、模式辨别与模式分析，现阶段使用较为普遍的 Web 数据挖掘技术是分类、聚类与统计分析等。

5.2 不同爬虫算法的比较

现阶段应用较为普遍的网络爬虫算法主要有广度与深度优先策略、Opic 策略以及 Partial PageRank 策略等，不同的网络爬虫算法的方式差异较大，每种算法都有着自身独到的优势，在具体应用过程中必须要联系应用场景科学选取网络爬虫算法方式。

5.2.1 广度优先策略

广度优先算法方式是严格按照 Web 内容各个目录级进行的，首先是对初始界面同一层级的页面进行爬取，紧接着将得到的链接依照相应的规律排列到队列当中，以此来达到向外拓展的目的，尽量得到更多的链接信息，同时继续向下一个层级深入。广度优先策略能够在同一时间对多个层级进行爬虫，从而有效改善 Web 信息抓取效率。正因为如此，在当下多个网络爬虫算法当中广度优先策略有着非常大的应用范围。然而该算法也存在一定的弊端，即当需要对一个拥有较多层次的目录进行挖掘时，通常要消耗非常多的时间，难以获得较高的挖掘效率。

5.2.2 深度优先策略

从字面意思来理解，深度优先策略是指爬虫以某个特定的顺序依次访问各个页面，并确保每个页面都挖掘到最底层的目录，当一个分支挖掘完成之后，才会退出并进行下一个分支的爬取，当对每一个链接的内容都爬取后，爬虫任务才全部完成。采用这种爬虫算法策略，可以实现对深层次信息内容的挖掘，然而当在对这些较深的站点内容进行爬取时，往往需要占据较多的系统资源。

5.2.3 Partial PageRank 策略

Partial PageRank 网络爬虫算法首先需要获取初始页信息，在此基础之上对 Web 层面的 Partial PageRank 的数值进行核算，并基于计算结果确定页面内在的价值大小，之后依照PageRank数值从大到小的方式依次开展各个页面的爬取，以此来有效改善完了爬虫开展的速度，同时还可以获得较好的遍历效果。然而这种网络爬虫算法也存在一定的不足，即

最终获得的爬取结果和实际的遍历结果有着较大区别，无法确保数据的精准性与可靠性。

5.2.4 Opic 策略

Opic 策略可以看成是在 Partial PageRank 策略基础之上的改进策略，在爬取准备阶段中，所有页面的数值均是一样的，当所有页面都完成下载任务以后，最高值将会平均配置给页面中的各个链接，于此同时初始化当前页面的数值，爬虫基于数值的高低对各个页面开展优先级排列，基于从大到小的原则进行页面的下载。Opic 策略不用进行迭代计算，能够很好的满足实际的计算要求 [12]。

5.2.5 数据结构化存储

一般来说，现有的信息当中绝大多数均是基于非结构的文本出现的，在实际进行信息的归类与应用过程中，都面临较大的难题，而将信息转变为结构化方式存储起来便能很好的解决上述问题。具体来说，首先从 Web 页面众多信息中将各个非结构信息单独挖掘出来，将其转变为结构化数据以后保存到计算机磁盘当中，以此来提高数据存储的规范性与标准性。值得注意的是，整个过程均可以实现自动化运行，工作人员不必进行相应的操作，使用者可以联系具体使用场景灵活选取数据库、CSV 等与之相匹配的存储途径。当确定选择数据库进行数据的存储时，Web 数据便会依托于二维表架构的方式完成各项数据的保存，这种数据存储方式不仅效率高，同时还拥有较高的精准性，可以很好的契合多线程数据挖掘的要求。

5.2.6 正则表达方式

事实上，绝大部分 Web 页面均是依托于 Html 格式呈现的，但是 Html 页面则是基于各种类型语义对象的基础之上，当对象不同时其标记也存在较大差异，当针对 Html 页面开展深入分析与探究后，结合正则表达方式进行配置，便可以有效查找与获取需要的字符串信息。例如，当希望获取到 Web 页面中涵盖“is”的程序代码时，需要利用 matchObj = re.match(r'(.*) is (.*?) .*', line, re.M|re.I) 的途径挖掘相关信息，以此来达到对涵盖“is”字符串的自主配置，同时输入相应的内容。在该示例当中，(r'(.*) is (.*?) .*', line, re.M|re.I) 便是为其配置的正则表达式，能够在数据挖掘过程中给予切实可行的方法。除此之外，为了能够妥善的解决页面健全、网站优化等问题，保障匹配过程的稳定性，使用者也能够选择 Python 自身拥有的模块和第三方数据库，实现对 Web 页面信息内容准确解析与获取。

6 结语

综上所述，本文重点阐述了 Python 语言的特征及其在 Web 数据挖掘中的应用优势，总结了网络爬虫算法优势与适用环境，同时对正则表达式、数据存储方式以及数据信息抓取过程进行了深入分析，得出 Python 语言可以很好的满足网络爬虫的数据抓取要求，可以进行数据的自主性、差异化抓取，有效改善了数据查找和分析的速度和质量。

参考文献

[1] 聂晶 . Python 在大数据挖掘和分析中的应用优势 [J]. 广西民族大学学报 (自然科学版), 2018, 24(01): 76-79.
[2] 柴文光，周宁 . 网络信息安全防范与 Web 数据挖掘技术的整合研究 [J]. 情报理论与实践 , 2009, 32(03): 97-101.
[3] 王小君 . 网络信息安全防范与 Web 数据挖掘系统的设计与研究 [J]. 电子设计工程 , 2018, 26(12): 83-87.
[4] 李彦 . 基于 Python 的数据挖掘——阳光集团的具体数据挖掘项目 [J]. 电脑知识与技术 , 2018, 14(23): 15-20+36.
[5] 王轶哲 . 基于数据挖掘的客户预测及其 Python 实现技术研究 [J]. 电子制作 , 2020(24): 51-52.
[6] 李宁 . 基于 python 的数据挖掘技术在公安情报分析中的应用研究 [J]. 电子世界 , 2020(06): 181-182.
[7] 卢林竹，王智浩，蒋益兰，何兰，曹如柔，蒋盛昶 . 基于 Python 语言构建名中医医案数据挖掘平台 [J]. 世界科学技术 - 中医药现代化 , 2021, 23(09): 3188-3194.
[8] 曾涛，阮彬 . 大数据挖掘与分析在项目管理中的机遇与应用——以 Python 技术为例 [J]. 中国管理信息化 , 2020, 23(23): 115-117.
[9] 王越，陈国兵，李军 . 基于数据挖掘的故障模式、影响及危害性分析改进方法 [J]. 科学技术与工程 , 2021, 21(24): 10536-10542.
[10] 张红军，王豫鑫，杨万里，祁永钊，李登明 . 基于大数据的数据挖掘中容错技术研究 [J]. 电脑知识与技术 , 2020, 16(09): 16-18.
[11] 王月梅，何雄伟 . 基于 Map/Reduce 的改进选择算法在 Web 数据挖掘中的研究与应用 [J]. 电脑知识与技术 , 2018, 14(23): 28-30.
[12] 郭宝琳 . 基于 Python 数据挖掘的进口服装质量语义词典构建研究 [J]. 质量技术监督研究 , 2020(05): 49-52.

作者简介

聂莉娟（1980-），女，江西省高安市人。硕士学位，讲师。研究方向为大数据。
方志伟（1982-），男，浙江省台州市人。硕士学位，讲师。研究方向为计算机网络。
赵心宇（1978-），男，河北省张家口市人。硕士学位，工程师。研究方向为计算机安全。