文章编号：2096-1472(2019)-08-21-03
DOI:10.19644/j.cnki.issn2096-1472.2019.08.007

基于python的WEB数据挖掘技术实现与研究

齐 慧
(山东科技职业学院信息工程系，山东 潍坊 261053)

摘 要：文章首先对web数据挖掘技术进行概括，分别从数据挖掘技术概念、技术应用优势与技术原理三方面进行论述。其次，重点探讨基于python基础上的web数据挖掘技术开发设计方法，对数据挖掘过程中的各类爬虫技术应用优势进行对比，可以作为数据挖掘系统构建过程中的理论参照。

关键词：python语言；数据挖掘技术；仿真实验
中图分类号：TP309 **文献标识码**：A

Research and Implementation of WEB Data Mining Technology Based on Python

QI Hui
(*Shandong Vocational College of Science and Technology*,*Department of Information Engineering*,*Weifang* 261053,*China*)

Abstract: This paper firstly summarizes web data mining technology,discussing the concept,application advantages and principles of data mining technology.Secondly,it mainly discusses the development and design methods of web data mining technology based on python,and compares the application advantages of various crawler technologies in the process of data mining,which can be used as a theoretical reference in the construction of data mining system.

Keywords: Python;data mining technology;simulation experiments

1 引言(Introduction)

运用web数据挖掘技术，能够模拟出用户基于网络环境中的浏览过程，并根据用户操作过程中的使用功能需求，自动跳转至指定的信息页面。通过数据挖掘，将无序并且数量庞大的信息自动提取存储，将其整理成为结构化的信息形式[1]。一方面，方便用户在信息浏览过程中对自身需要的数据进行存储，另一方面也能够根据数据挖掘对各类功能进行表达，满足用户信息浏览过程中的不同需求。数据挖掘技术使用范围十分广阔，能够用于不同区域，并且在功能整合过程中也能够根据最终的综合控制能力，判断接下来的数据挖掘方向[2]。数据挖掘技术是存储功能实现不可缺少的基础，也具有极强的整合能力，能够与其他技术方法相结合，高效便捷的完成数据捕捉和存储。数据挖掘技术在不同领域均充当着重要角色，将web数据挖掘技术，与学习型汇编语言相结合，在程序设计过程中更能够体现出人性化功能，也能处于网络环境下，对数据信息进行高效定位，实现安全便捷的数据挖掘以及功能指令传输。

2 数据挖掘技术发展优势(Development advantages of data mining technology)

随着网络信息技术不断发展进步，数据挖掘技术也具有广阔的应用前景。网络环境中的各类数据信息资源，并没有固定结构存在。浏览网络信息中对于其中的有用数据提取往往会消耗过多时间。通过互联网技术普及，数据挖掘技术的应用能够将零散的信息进行整合，并根据用户不同使用功能选择自动或手动的挖掘存储[3]。数据挖掘技术在信息整合速度上十分快，具有极强的技术适应能力，应用该技术能够体现出不同挖掘项目之间的统筹能力，并根据挖掘过程中体现出的多角度问题[4]。新型技术应用方向调整，基于python语言基础上的网络系统设置，能够明显降低数据传输过程中的误差，并帮助查找遗漏，对遗漏数据自动填补。尤其是面对统计任务量较大的数据时，能够快速完成信息分类对接，并根据用户使用过程中的各类规则，对程序进行调整，纠正程序中存在的错误。数据挖掘过程中，能够确定数据传输的最佳路径，从而在传输过程中节省时间。由此可见，数据挖掘技术具有明显的发展优势，未来技术发展中，也将进入到更理想的状态中，通过不同汇编语言之间的相互结合，达到理想的设计效果。

3 数据挖掘技术应用原理(Application principle of data mining technology)

数据挖掘技术在应用过程中的功能实现，通过对用户基

于网络环境中浏览信息的脚本捕捉，自动进行有效数据信息排序，并根据用户所发出的功能指令对有用信息进行子集合构建，并对信息系统中的数据进行访问。访问wed页面内的相关内容后，根据反复的信息验证。数据挖掘功能原理如图1所示。

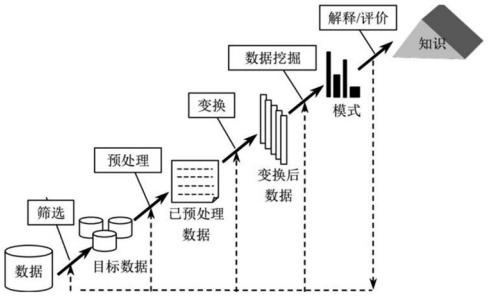

图1 数据挖掘原理

Fig.1 Data mining principle

确定最终的有用信息结合范围，从而实现子页面内的各类数据结合，进行切实有效的数据信息分类与整合。信息分类整合同样是实现模块化管理的基础，也是数据挖掘功能实现的原理，在数据挖掘过程中，会涉及不同爬虫算法的使用，选择的爬虫算法直接关系到数据信息挖掘，提取速度与最终的数据集合构成稳定性。数据挖掘技术在功能实现方面，需要对不同爬虫算法进行对比，从用户登录web页面后的起始页至最后一页进行连续的数据提取逐渐向外层延伸，并构建多角度信息获取链接，自动实现信息的捕捉[5]。数据挖掘与数据提取是相对应的功能，挖掘后并确定数据的来源范围，才能进行下一步功能构建。提取数据后并将其发送到指定的功能层，在页面功能实现过程中，筛选有用信息并进行结构化整合，经过数据搜索与分析最终确定挖掘对象，实现一系列数据提取功能。

4 数据挖掘技术中的算法比较(Comparison of algorithms in data mining)

4.1 广度优先算法

数据挖掘技术应用过程中，算法的比较研究内容比较多，首先是广度优先的算法策略，在计算过程中从起始页到最终的页面，要进行由内而外的延伸运算。并对多链接信息进行整合，在数据挖掘过程中自动进入到下一集层的深度中，确保数据挖掘在web网络环境中的广度。在挖掘分析过程中，对不同目录进行深入分析，确保挖掘过程中的分析内容涵盖整体目录。其优势在于广度优先策略，在运算过程中精准度十分高，其劣势在于挖掘过程中对目录分析将会耗费大量时间。广度优先算法主要是针对目录精准排查，实现链接的提取与扣件。能够进行算法的并行处理，同时在Web信息的挖掘，提取出多少也会有所提升。如果挖掘数据信息涉及到深层目录，最终的功能将会受到影响。

4.2 深度优先算法

深度优先算法应用在数据挖掘技术中，注重在同一区域范围内的深层次数据捕捉。根据用户的浏览内容在当前页面访问时，会进行深层次数据挖掘，直到在当前页面的最深点数据挖掘成功后，视为完成一个分支任务。并返回到最初的访问界面，从而进入到另一个爬行分支中进行相关数据的挖掘整理，直到对所有链接的深层次分析结束后，完成整体爬行任务。算法流程语言如下：

```
import re
from bs4 import BeautifulSoup
print bsObj。prettify
urls=soup.findAll(”a”,hrel=True)
defgetLink(countryUrl);
html=urlopen(' https：//baike.baidu.com/'+itemUrl)
bsObj=BeautifulSoup(html,'html.parser')
returnbsObj.finall(“a”,href=re.compie(“”(/item/)((?!：).)* ba
sicInfo-item value”))
links=getLinks(“https：//baike.baidu.com/item/%E5% 9B%
BD%E5%AE%B6/17205”)
whilelen(links)>0
links=getLinks(newCountry)
for link in bsObj.findALL(“a”)
if “href” in link.attrs；
print(links.attrs[‘href’]
```

该种分析方法，能够确保挖掘信息的深度，但如果在挖掘过程中，需要对更深层次的数据进行捕捉，将会消耗大量的分析资源。深度优先算法对于低层站点的数据挖掘和统计，这种效果并不理想，并且在最终的数据对比分析中，容易在某一链接范围内产生误差。因此该种技术手段应用，还需要进行技术方法之间的相互结合，达到最佳控制效果。

4.3 数据结构化存储

数据结构化存储也是数据挖掘过程中最常使用的技术手段，结构性存储能够针对原本杂乱无序的数据信息进行归类整合，并达到最佳的结构化存储形式。通过无结构信息的提取，并将其整合成为另一种链接形式，存储到本地文档中。能够确保数据信息的存储形式得到规范统一，并在执行过程通过人工整合达到最理想的场景构建模式。在存储过程中，结构化处理需要确保准确度与速度，既要满足多链接数据挖掘需求，同时也能够根据存储结构的调整，快速实现各链接之间的相互结合。结构化存储功能对于数据的综合处理能力十分快，处于Web环境下能够实现数据信息的自动结构调整，并通过结构之间的相互转换，减少人工操作带来的数据误差，结构图见图2。

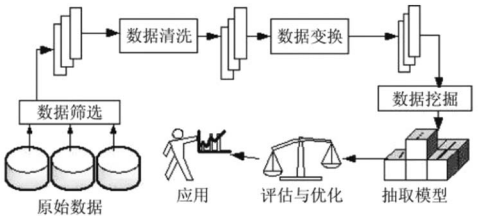

图2 数据结构优化图

Fig.2 Data structure optimization diagram

结构仅仅作为数据存储中的一种形式，在统一结构的同时，还需要考虑不同数据信息类型。自动选取最为高效便捷的存储方式，保证数据信息在存储过程中的安全性与使用效率。数据结构化存储对数据挖掘技术精准度要求极高，不仅在稳定性与速度上要达到使用标准同时还需要满足自动归类功能，能够在归类过程中增强综合控制能力，实现数据结构化的自动存储，在结构化存储过程中自动生成二维表格，达到最佳功能整合效果。

5 基于python的WEB数据挖掘设计(Design of WEB data mining based on python)

5.1 爬虫功能设计

基于python语言技术基础上的web数据挖掘设计，首先需要对爬虫功能进行选择，根据用户日常使用过程中对于功能的特殊性需求，对比不同爬虫方法之间的优势与劣势。借助python技术的分析功能，在计算过程中对数据的广度进行扩增，并根据不同数据以及关键词在网络信息浏览中的出现次数，进行自动分析定位，确定接下来的语言扩增形式。数据结构设计过程中，不仅需要对稳定性进行对比，还需要根据数据抓取过程中的链接分析，进行最终的匹配链条确定。爬虫功能根据浏览页面的实际情况，对脚本内容进行构建，通过应答服务体系以及构建过程中的超链接获取，实现对数据信息的快速筛选。但在最终的数据信息获取和整合过程中，根据所分析的内容进行最终的数据整合。并在挖掘过程中对所涉及的各类功能进一步调整，针对数据挖掘设计中的功能在强化过程中体现出多元化整合能力。数据分析时对页面的源代码进行提取，在源代码分析基础上进入到更稳定的数据整合阶段，并根据场合得到的各类结果，采取多元化调整措施，提升数据信息之间的相互配合能力。

5.2 数据表达设计

数据表达设计过程中，一方面要考虑数据挖掘与最终使用的稳定情况，另一方面也需要根据数据的具体表达能力。在设计过程中体现出最佳的表达方法，对数据的构建形式加以完善。数据表达中需要考虑不同页面访问的过程调整，并根据元数据体系最终的判断，在表达形式上体现出控制指令之间的对接能力。对于数据表达过程中不同方法理念之间的选择以及构建，更需要多元化的融合角度体现出数据表达的综合控制能力，尤其是在数据表达设计阶段，各个功能方法之间的相互结合，充分体现出元数据的多元化控制能力，以及最终的数据综合挖掘情况。在不同功能页面，采用多种结构的形式对数据进行表达，数据表达后才能进入到接下来的有用信息捕捉与自动存储阶段。数据表达设计阶段，同样需要借助python语言来进行模拟设计，实现数据表达过程中的爬虫功能，以及在数据挖掘提取阶段不同功能之间的相互控制能力。运用多元化数据整合模式，进行分层结构完善以及结构化功能的实现，完成数据挖掘、数据提取和结构构建多元化功能之间的融合。数据表达设计期间的综合控制能力提升，以及最终的运行状态调整，还需要在管理阶段体现出数据的综合表达能力，对表达流程和表达形式进一步设计，实现数据表达与数据提取一体化模式。

5.3 仿真功能检验

仿真功能检验是指在功能应用过程中，对于所构建设计的全部系统以及数据信息提取形式进行仿真功能验证，观察是否在仿真功能上能够达到预期效果，以及最终的仿真能力是否与开发设计中所确定的功能目标保持一致。对于开发设计阶段所确定的各类功能，需要通过仿真实验后确定其可行性，才能在接下来的系统中制定进一步的综合控制目标。仿真实验需要模拟网络环境中潜藏的风险隐患，对所构建的系统结构进一步整合，观察系统结构的综合控制能力。模拟病毒对系统进行攻击，从而判断系统开发设计中需要进一步完善的内容。仿真功能检验过程中，对于所存在的问题，需要将其划分到同一集合中。脚本仿真实验程序如下：

```
fromurllib.request import urlopen
from bs4 import BeautifulSoup
html=urlopen(' https：//baike.baidu.com/')
bsObj=BeautifulSoup(html,'html.parser')
t1=bsObj.find_all('a')
for t2 in t1：
t3=t2.get('href')
print(t3)
```

6 结论(Conclusion)

在接下来的开发设计阶段，重点针对现存问题部分加以完善，并通过仿真功能检验对问题进行拓展，观察是否存在系统之间的相互影响。并通过干扰分析增强最终的仿真功能稳定性，根据所得到的仿真功能检验结果，确定系统在网络环境中运行使用的薄弱环节，通过数据挖掘技术，增强最终的功能稳定性。

参考文献(References)

[1] 王雪峰.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.

[2] 邢娜.浅析基于Web数据挖掘应用于电气自动化技术对社会经济发展促进作用的研究[J].青春岁月,2017(12):427.

[3] 李岩松.集成Vissim和Python的车联网仿真平台研究[J].计算机仿真,2018,35(12):159-162;421.

[4] 唐琳.基于Python的自然语言数据处理系统的设计与实现[J].电子技术与软件工程,2018,138(16):176-178.

[5] 黄雪华.基于Python的决策树算法在学生招生录取数据中的应用研究[J].电脑知识与技术,2018,14(29):22-23.

作者简介：

齐　慧(1982-)，女，本科，助教.研究领域：云运维，web前端.