基于 Python 聚焦网络爬虫的用户在线评论内容分析

作者：王煜炜

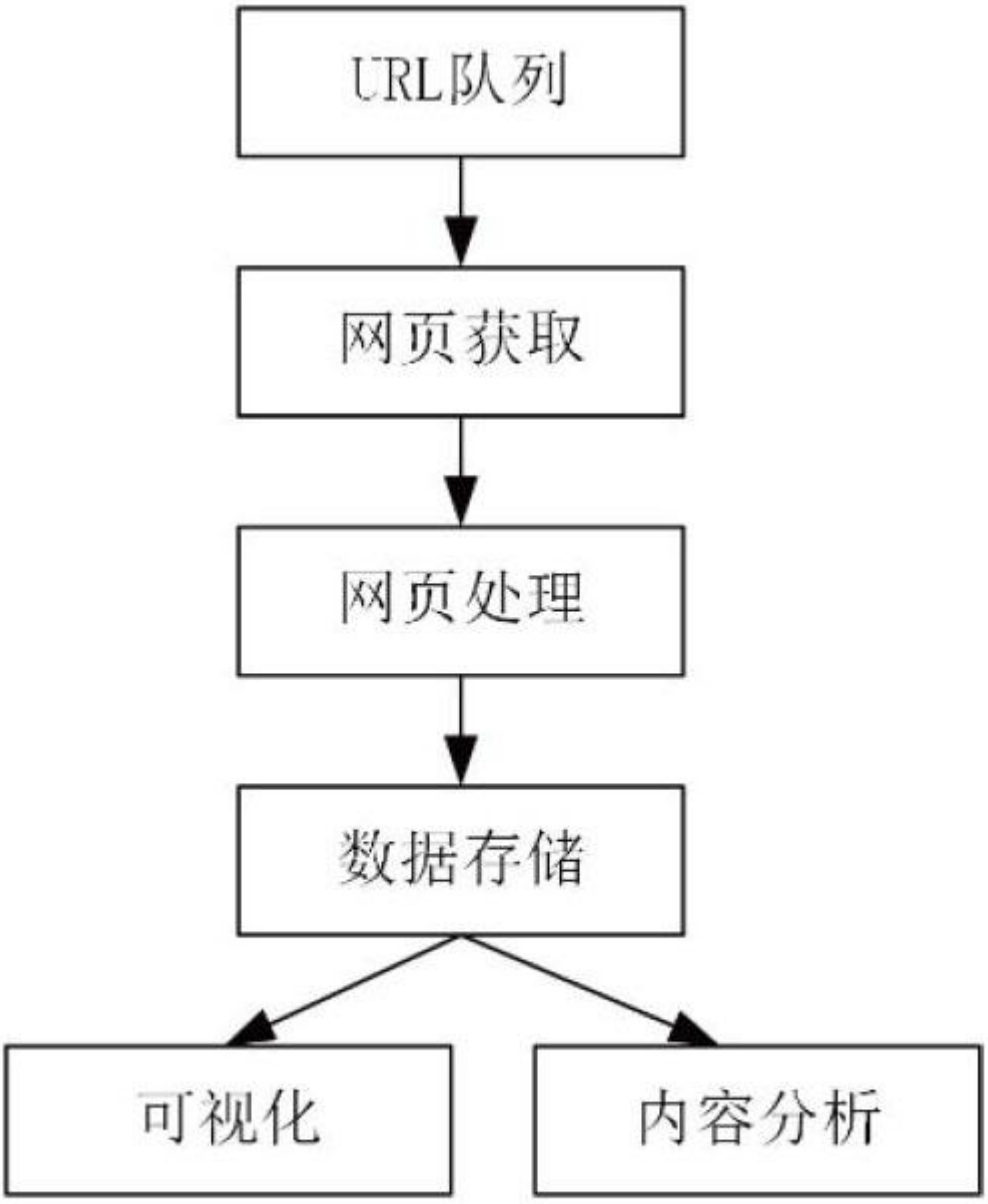

图1　聚焦网络爬虫系统工作流程图

```
{
    "comment_id":"140572",
    "business_id":"14",
    "user_id":"948420",
    "user_name":"214137",
    "comment_rank":"1",
    "content":"场地不好，又热，人不多",
    "create_time":"1500535753",
    "avatar":"http://api.7yundong.cn/uploads/avatar/avatar.png",
    "image_list":Array[0]
}
```


图2　用户评论信息JSON结构解析实例

图3　用户在线评论信息词云图

摘 要 近年来，随着“互联网+”的全面推进，互联网上的信息量不断增加，如何高效、快速地获取目标信息，并对信息进行有效分析成为亟待解决的问题。本研究设计并实现了一个基于Python 的聚焦网络爬虫系统，以体育场馆用户在线评论为例，对评论信息进行获取，并对数据进行可视化展示和内容分析，结果表明，该系统能够较好挖掘用户对于场馆评论中隐藏的信息。

关键词 Python;聚焦网络爬虫;在线评论;内容分析

引言

物联网、人工智能、大数据、5G 等技术的不断发展和成熟，有效促进了互联网数据的增长。互联网数据通常具有海量、多维、多尺度等大数据的特点，采取有效的方法和手段对互联网数据进行收集和分析，是理解互联网数据的重要手段[1]。传统的数据收集方法和技术（如调查问卷，采访）会受到资金、地理位置和样本量等一系列条件的影响和限制。网络爬虫依托于大数据框架和计算机技术，可用于对海量互联网数据进行请求和提取，为深层次的内容分析和挖掘奠定了基础[2]。

本文基于 Python 标准库、第三方库和爬虫技术，设计并实现了一个聚焦网络爬虫系统，用于对指定网页和内容进行爬取，并进一步对爬取到的数据进行存储和内容分析。在实例分析中，爬虫系统第一步将抓取“趣运动”网站（http：//www.quyundong.com/）中体育场馆用户评论信息的页面，第二步对网页进行解析并过滤无关的内容和数据，第三步对抓取数据进行存储，第四步对体育场馆用户中差评信息进行词云图展示和词频统计，第五步对高频词汇和词云图进行分析，挖掘造成中差评的主要原因，为场馆改进提供建议，同时为其他用户提供决策依据。

1 基于 Python 的聚焦网络爬虫系统设计

1.1 聚焦网络爬虫定义

通用网络爬虫通过统一资源定位符（Uniform Resource Locator，URL）搜索网页，通过遍历所有待抓取 URL 队列，将网页相关数据返回给用户[3]。聚焦网络爬虫基于通用网络爬虫，专注于抓取满足特定主题和特定属性的网页。该爬虫策略性搜索、获取、下载、维护与特定主题相关的网页 URL，所有其他无关的 URL 将通过程序代码被过滤。通过采用聚焦網络爬虫，用户无须通过网页搜索引擎来获取信息，这样既节省了时间和精力，又提高了数据采集的可靠性、针对性和准确性[4-6]。聚焦网络爬虫下载的“面向主题”的数据，后续可通过采用有效的内容分析和挖掘技术，提取出数据中隐藏的有价值信息。

1.2 聚焦网络爬虫系统工作流程

该聚焦网络爬虫系统工作流程分为下列 5 部分，如图 1 所示。

（1）URL 队列：聚焦网络爬虫系统基于指定一个或几个网页网址，把这些网址作为 URL 种子，将 URL 种子放入 URL 队列中等待爬取。

（2）网页获取：根据指定 URL，按照一定的规则对网页进行遍历，发送请求并执行相应爬取。

（3）网页处理：对网页信息进行解析和处理，提取出与研究主题相关的网页内容部分，过滤掉其他无关数据和内容。

（4）数据存储：对进行网页处理后，与研究主题相关的数据进行存储，本研究中将其存储为 Excel 格式。

（5）可视化和内容分析：对存储数据进行可视化以及内容分析，包括词云可视化分析、词频统计等方法。

2 数据爬取与存储具体实现

本章借助 Python 聚焦网络爬虫系统，以爬取 “趣运动”网站体育场馆用户在线评论为例，进行数据爬取与存储的具体实现，分为以下 3 步。

（1）网页抓取。趣运动网站采用的是异步加载 Ajax 技术，通过分析趣运动网站结构和网址构造，得到获取用户评论的 URL 请求地址为 http：//www.quyundong.com/venues/jsonComments？random=xxxx&page=****&business_id=####，该请求由 3 个网页参数组成：xxxx 对应的是随网页请求生成的随机数（random）、****对应的是在线评论页数（page）、####对应的是场馆编号（business_id）。确定 3 个参数后，采用 Python 中的第三方 requests 库对指定场馆用户在线评论 URL 请求进行抓取。

（2）网页处理。趣运动网站用户评论信息以 JSON 格式进行存储，评论信息的 JSON 结构如图 2 所示，故调用 Python 中的 JSON 库对评论信息进行解析。由于评论信息以键值对的形式存在，在遍历每位已注册用户评论信息时，仅需筛选出评论时间（create_time）、评论内容（content）、评论等级（comment_rank），其他的信息：评论 ID（comment_id）、场馆 ID（business_id）、用户 ID（user_id）、用户姓名（user_name）、用户头像（avatar）、用户上传图片列表（image_list），由于与该研究主题关联不大，将被过滤掉，不参与数据爬取。最终将所有符合要求的评论信息存入一个结果集中。

（3）数据存储。调用 Python 中的 Workbook 库，将第二步得到的结果集写入 Excel 文件，对体育场馆用户评论数据进行存储。最终获取到用户有效在线评论数据 18023 条，从这些数据中筛选出评论为 3 分及以下（comment_rank<=3，满分 5 分）的中差评共 768 条，作为可视化和内容分析的对象。

3 可视化与内容分析具体实现

在聚焦网络爬虫系统架构中，爬取完所需数据之后，需对数据进行可视化和内容分析，旨在挖掘出数据中隐藏的有价值信息。

（1）数据可视化。读取所有用户在线评论文本，导入 Python 中的 jieba 中文分词库，获取在线评论的中文分词列表。接着使用 wordcloud 词云库，设置 stopwords 屏蔽词参数，对数据进行清洗，同时设置词云图的形状、背景颜色、高度、宽度和字体，结果可生成相应词云图，对场馆用户评论数据中出现频率较高的“关键词”予以可视化的展示（如图 3 所示）。接着调用 Sklearn 库中的 CountVectorizer 函数，分别提取词汇和计算词频，对评论数据中的词汇进行词频统计，并将结果存储在 CSV 文件中。

（2）内容分析。对筛选出的评论为 3 分及以下的中差评数据进行统计，其中评分为 3 分的用户评论共 389 条，占比 50.65%;评分为 2 分的 106 条，占比 13.80%;评分为 1 分的 260 条，占比 33.86%;评分为 0 分的 13 条，占比 1. 69%。

结合词云图和词频分析结果，发现出现次数最多的前 10 个高频词汇分别是：场馆（286 次）、不好（130 次）、灯光（84 次）、服务态度（70 次）、没有（68 次）、价格（65 次）、地板（54 次）、空调（54 次）、位置（49 次）、态度（38 次）。由此可以得出以下结论，用户对于使用体育场馆的需求主要包括：场馆的灯光、收费价格、地板、空调、位置和工作人员的服务态度。该结果基于聚焦网络爬虫系统数据收集、存储、分析整个流程，反映了趣运动网站用户在线评论的真实情况，挖掘了用户在参与体育场馆设施健身过程中的具体需求。

4 结束语

本文基于 Python 构建了聚焦网络爬虫系统，实现了对趣运动网站体育场馆用户在线评论信息的爬取、存储和内容分析，分析结果表明：聚焦网络爬虫专注于特定主题和内容的收集，提高了信息收集效率，节省了大量的时间。同时，对用户评论数据的存储便于进一步的数据管理和分析，也能对有效数据进行保存。再者，对存储数据的可视化和内容分析结果，包括词云图可视化和词频统计，可进一步为场馆硬件和软件设施改进提供决策，也可为用户选择体育场馆提供参考。上述结论验证了本文提出的聚焦网络爬虫系统的有效性和实用性，下一步的工作是继续优化聚焦网络爬虫系统，结合多线程与并发等技术，同时加入更多的文本分析算法，对整个系统性能进行优化，使其爬取效率更高、功能更加完善。

参考文献

[1] 杜晓旭，贾小云.基于 Python 的新浪微博爬虫分析[J].软件，2019， 40（4）：182-185.

[2] 刘晖，石倩.基于网络爬虫的新闻网站自动生成系统的设计与实现[J].电子技术与软件工程，2019（13）：18-19.

[3] 陆树芬.基于 Python 對网络爬虫系统的设计与实现[J].电脑编程技巧与维护，2019（2）：26-27，51.

[4] 郭向向，郑嘉慧，苗学芹.基于 Python 聚焦型网络爬虫的影评获取技术[J].时代金融，2019（11）：71-72.

[5] 高宇，杨小兵.基于聚焦型网络爬虫的影评获取技术[J].中国计量大学学报，2018，29（3）：299-303.

[6] 杨国志，江业峰.基于 python 的聚焦网络爬虫数据采集系统设计与实现[J].科学技术创新，2018（27）：73-74.

作者简介

王煜炜（1990-），男，湖北武汉人;毕业院校：武汉大学，专业：软件工程，学历：博士研究生，现就职单位：江汉大学，研究方向：人工智能。