文章编号：2095-2163(2019)04-0112-05 中图分类号：TP311 文献标志码：A

校内综合信息服务平台关键技术研究与实现

龚 丹，石蕴金
（哈尔滨华德学院 电子与信息工程学院，哈尔滨 150025）

摘 要：当前在大学校园中已经建设的各种信息系统互相独立、分散性很大，已不能满足移动互联网发展过程中，人们对信息获取便捷性、快速性的需求。本文首先简述了网络爬虫和 OCR 验证码识别技术；然后结合统一文件存储、外部数据库代理、Docker 虚拟化容器等云平台、跨平台技术，给出接入已有校内系统获取信息，进而建设一站式校园信息综合服务平台的解决方案；最后给出方案实施后在功能、性能和实用性等角度的测试结果。结果表明本方案可以打破传统校内系统的壁垒，向用户提供更便捷、移动端友好的信息服务。

关键词：校园信息系统；网络爬虫；OCR；云平台

Research of key technologies and implementation of campus integrated information service platform

GONG Dan, SHI Yunjin
(School of Electronic and Information Engineering, Harbin Huade University, Harbin 150025, China)

【Abstract】 At present, a variety of information systems have been built on university campuses, which are independent of each other and decentralized. So far, it can not meet the need of convenience and rapidity of information access, especially in the current era of mobile Internet. In this paper, firstly, the key technologies of network crawler and OCR verification code recognition are studied. Then, the solution of building a one-stop campus information integrated service platform is given, which combines some cloud platforms and cross-platform technologies, such as unified file storage, external database agent, Docker virtualization container, etc. Finally, the test results of the scheme in the aspects of function, performance and practicability are given. The results show that this scheme can break down the barriers of traditional campus information systems and provide users with more convenient and mobile friendly information services.

【Key words】 school information system; Web crawler; Optical Character Recognition; cloud platform

0 引 言

当前大学校园通常都配备有教务系统、图书馆系统、校园卡系统、门禁系统、就业系统等等。这些系统覆盖了高校内的各种核心业务，但是通常是各自为政[1-3]。信息化管理的普及给工作带来高效的同时，也积累了越来越多的应用系统；尤其是移动互联网发展，移动端 APP 疯狂上市，近期央视新闻报道在一些高校信息化手段出现的过度倾向[4]，打水、连网、记学分都要 APP。可见，当前信息化手段的应用不再局限于管理工作，更多地是提供服务，以方便师生在校学习、工作时，方便快捷地获得所需要的信息。这种信息获取的需求，内容是基础，但更重要的是体验——随时随地、快速便捷、清晰美观地呈现给用户。因此，作者以工作和学习的校园实际情况为背景，提出打破已有独立系统壁垒、提供一站式综合信息服务系统的方案，并利用云平台技术，去除用户安装和使用本系统的计算资源负担，享受高品质的信息服务。

1 相关技术

1.1 网络爬虫

网络爬虫是一个自动地从互联网上抽取网络信息的程序，通常作为网络数据收集的工具。通过编写爬虫程序，完成特定的过程和算法，从目标地址中获取所需要的信息，并使用特定算法完整数据结构分析和整理[5]。一般爬虫分为增量型爬虫、批量型爬虫和垂直型爬虫。增量型爬虫无固定范围目标，持续不断的抓取互联网中的各种类型的信息，根据目前互联网的变化而不断变化抓取的内容。批量型爬虫有固定范围目标，设定一定的目标达到设定的目标就自动停止抓取信息。而垂直型爬虫不像通用

作者简介：龚 丹(1979-)，女，硕士，副教授，主要研究方向：计算机应用技术、可靠性软件工程；石蕴金(1998-)，男，本科生，主要研究方向：网络工程。
通讯作者：龚 丹 Email：gondan1979@hotmail.com

收稿日期：2019-03-18

爬虫那样需要全面地从互联网抓取网络数据，而只编写成为完成特定目的，为指定系统和目标使用特定算法抓取目标信息的爬虫[6]。本文仅研究垂直型爬虫以实现获取系统所需的信息和功能。

1.2　OCR 验证码识别

1.2.1　OCR 技术介绍

OCR 技术是光学字符识别的缩写（Optical Character Recognition），通过扫描等光学输入方式将各种票据、报刊、书籍、文稿及其印刷品的文字转化为图像信息，再利用文字识别技术将图像信息转化为可使用的计算机输入技术。可应用于银行票据、大量文字资料、档案卷宗、文案的录入和处理领域。适合于银行、税务等行业大量票据表格的自动扫描识别及长期存储。相对一般文本，通常以最终识别率、识别速度、版面理解正确率及版面还原满意度 4 个方面作为 OCR 技术的评测依据；而相对于表格及票据，通常以识别率或整张通过率及识别速度为测定 OCR 技术的实用标准[7]。

1.2.2　验证码识别的概念

验证码的英文 CAPTCHA，即全自动区分计算机和人类的图灵测试，是为区别对方到底是人类还是计算机程序而设置的一种验证措施，主要用来防止网络机器人的一些恶意行为[8]。

验证码是一种区分用户是计算机和人类的全自动程序。该程序生产一个问题，可以由计算机生成并评判，必须只有人类才能解答该问题。由于计算机无法解答验证码的问题，所以回答出这个问题的用户就能被认定为是人类。由于这个测试是由计算机来考人类，而不是标准的图灵测试中那样由人类来考计算机，被称为反向图灵测试。

2　网络爬虫的详细设计

2.1　正方教务系统爬虫

正方现代教学管理系统是目前广泛用于高校学院各部门以及各层次用户的多模块综合信息管理系统，包括教务公共信息维护、学生管理、师资管理、教学计划管理、智能排课、考试管理、选课管理、成绩管理、教材管理、实践管理、收费管理、教学质量评价、毕业生管理、体育管理、实验室管理以及学生综合信息查询、教师网上成绩录入等模块。正方教务系统爬虫是针对正方教务系统自动获取系统中的学生数据、成绩数据、课表数据等数据并通过特定算法完成整理数据结构返回其数据信息的垂直型爬虫。具体结构如图 1 所示。正方教务爬虫的基础流程如下：

（1）建立抓取指定信息的任务；

（2）判断缓存数据库是否含有任务缓存信息；

（3）尝试进行模拟登入；

（4）将网页信息和验证码交给统一文件存储服务；

（5）通过 OCR 验证码识别服务识别验证码；

（6）指定信息网页下载；

（7）根据数据格式要求调用指定算法进行数据结构整理；

（8）格式化数据并进行编码；

（9）更新存储数据到缓存数据库。

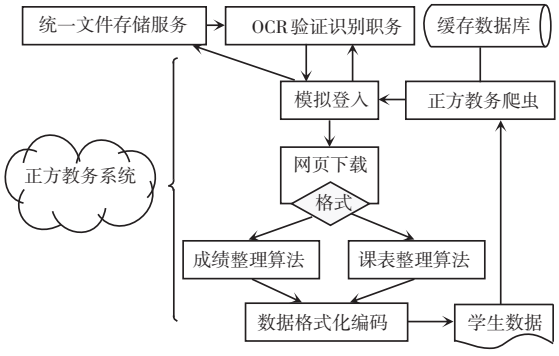


图 1　正方教务系统爬虫结构图

Fig. 1　Crawler framework of ZHENGFANG educational management system

2.2　锐捷网络管理系统爬虫

锐捷 RG-SAM 认证计费网络管理系统是目前广泛用于高校学院中的一种认证计费网络管理软件。锐捷网络管理系统爬虫是针对锐捷 RG-SAM 认证计费网络管理系统会根据用户的信息数据自动获取用户网络数据的用户余额、上网周期、绑定 IP、网络参数等信息的垂直型网络爬虫。具体结构如图 2 所示。锐捷网络管理系统爬虫的基础流程如下：

（1）建立指定信息抓取任务；

（2）系统登录；

（3）将信息存入统一文件存储服务；

（4）使用 OCR 验证码识别服务识别验证码；

（5）指定信息网页下载；

（6）使用数据整理算法；

（7）返回数据信息。

3　OCR 验证码识别服务的详细设计

验证码识别服务主要有以下几个主要功能，实现图像的采集功能、去噪、二值化、字符切割、样本的训练和识别。在本文中，作者对系统的总体功能结构、技术框架和总体类进行了分析和实现，在此基础上，本章将对验证码识别服务的主要功能进行详细

设计。

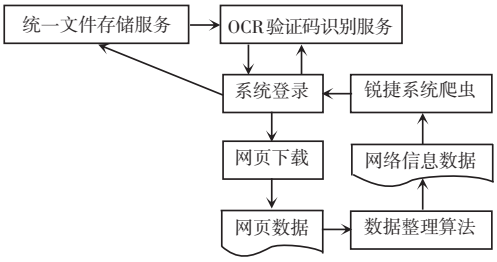


图 2 锐捷网络系统爬虫

Fig. 2 Crawler framework of RUIJIE network management system

3.1 图像采集功能

通过特定方式请求,从 Internet 上获取需要识别的验证码并存储在统一文件存储服务中,并将存储信息告知识别程序。程序流程:

(1)由服务端传输所需要的请求信息,如:请求 id 值、请求类型、请求地址、COOKIE、HEADER 等信息。由图像采集功能模块建立 socket 链接从 Internet 上读取图片文件;

(2)将读取到的图片文件流通过请求 id 命名、PUT 提交给统一文件存储服务进行存储;

(3)告知服务端图片获取是否成功,由服务端继续执行图片识别服务。

3.2 图像的去噪

验证码在生成中为了防止识别添加了许多噪点或传输压缩的过程中产生了很多噪声,去噪声是图像处理中常用的技术手段。通常需要对识别目标的噪声进行分析,根据不同噪声的特性进行去噪,也称噪声去除。

根据采集到的大量样本表明字体颜色采用了纯蓝色 RGB(0,0,153),使用该特征可以很方便地去噪。将图片转换为矩阵,进行循环扫描滤波,去除其颜色信息后,获得只包含字符信息的图片矩阵。

3.3 图像的二值化

为了让程序更好更快地识别其中的信息,需要对彩色图像信息进行处理,图像的二值化就是把图像中的像素根据一定的标准分化成 2 种颜色。区分图像中的前景信息与背景信息,简单定义前景信息为黑色,背景信息为白色。

3.4 图像的字符切割

一般图像信息中含有多个字符,识别时需要根据每个字符的特征进行比对识别,所以对图像进行字符切割是不可或缺的。这一步的主要工作就是把图像中的字符独立出来以方便识别和进行处理。

对图像信息分析发现,字符使用的字号相对固定,每个字符不会出现重叠现象仅可能出现粘连,文字旋转方向相对不固定但小于等于 45°。规则字符的粘连很容易分割开,如果字符本身有缩放、变形就很难处理。经分析,可以发现,上面的字符粘连属于很简单的方式,只是规则字符的粘连,处理这种情况,可使用很简单的处理方式。当完成分割操作后,不能马上确定分割的部分是否为一个字符,要进行验证。验证的关键因素就是,切割下来的字符的宽是否大于阈值,这个阈值的取舍标准是,一个字符无论怎么旋转变形都不会大于这个阈值。如果切割的块大于这个阈值,就认为这是一个粘连字符;如果大于 2 个阈值之和,就认为是三个字符粘连,以此类推。了解这个规则后,切割粘连字符也就很简单了。如发现是粘连字符块,直接平分这个块为 2 个或者多个新的块就可以。当然为了更好地还原字符,本文采用平分+1、-1 对字符块的部分进行适当的补充。具体算法如下:

(1)扫描整个图像信息,取得真实字符串图像所在位置和大小,进行记录;

(2)按照字符串数量进行等分,并判断字符是否存在粘连并进行处理,取得每个字符串所在的大概位置;

(3)将图像进行分割,取得每个字符图像。

3.5 图像的训练和识别

首先,采取大量训练样本来进行处理分析,取得每个字符大量样本后进行标注,将得到的信息以文件形式保存为权值矩阵。本程序中采用的训练样本以图片形式展现。经过处理后的训练样本仅保留了字符的特征信息,将大量样本用来扩充样本库。识别服务启动后首先载入全部样本库,识别样本时创建一个权重数组,按照相似度比对进行权重累加。最后输出权重值最高的字符作为识别结果进行返回。经过挑选采用了 800 个含有字符数据信息的图像作为训练样本。图像中包含了 0 到 8 的 9 个数字、不包括 O 的从 A 到 Y 的 24 个英文小写字母,总共 33 个字符处理后的图像。

经过测试,这些训练样本训练后的识别服务对于其验证码可以达到约 90%以上的识别率。当然如果进行图像倾斜度矫正,那无疑可以进一步提高识别率,但是图片处理的时间和复杂度会达到非常大的级别,同样如果再增加训练样本也可以提高识别率,所以实际意义不大。

4 方案实施与应用

除上述关键技术,本文系统的总体设计方案中

将系统按模块划分为控制节点模块、交付模块、自动发布模块和 Web 管理页面模块。采用 Docker 虚拟化容器架构，提高了跨平台的部署性、扩展性和安全性。系统分为以下子容器系统：Web 平台服务、服务注册系统、自动发布服务、容器管理服务、OCR 验证码识别服务、统一文件储存服务、外部数据库代理服务、数据库服务。最后，将系统部署到学校的服务器集群上，提供云平台计算，应用前端通过微信小程序发布（微信搜索小程序“华德校园”，试用账号请与作者联系），供用户使用。本部分给出对关键技术的测试和系统应用情况的展示。

4.1　信息抓取与格式化——以课程表信息为例

在校内各项信息中，课程表是学生最为关注的信息之一。因此，本部分以展示前面所设计的正方教务系统爬虫的实施效果为例，以说明本系统的信息服务功能。如图 3 所示，（a）为算法直接抓取到的数据，显然，需要经过格式化，才能成为课表服务中的输入，并以直观易懂的形式显示给用户。（b）为从抓取数据中提取出关键信息的中间表示形式。鉴于大学课程在不同教学周的安排变化较大，本平台最终输出给用户的课程表，以周为单位进行展示，如图 4 所示。

（a）原始数据　　（b）格式化后数据

（a）Raw data　　（b）Formatted data

图 3　课程表数据格式化对比

Fig. 3　Comparison of raw data and formatted data

4.2　验证码识别服务

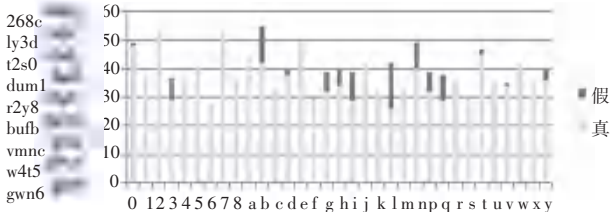
如图 5（a）所示，带阴影的图文区域是旧系统登录时随机生成的验证码，其左侧黑色字符为本文方法识别得到的结果。图 5（b）是对本方法进行 330 次测试得到的统计结果，该验证码系统中 33 种字符，识别准确率在 90% 以上的字符 24 种，占 73%；准确率在80%以上的字符则为28种，占85%；识别效果最差的为字符 1，准确率为 62%。本服务设置了自动尝试 3 次，因此验证通过率完全符合实际需要。

（a）APP 首页　　（b）“课程表”服务

（a）UI of APP　　（b）The curriculum schedule service

图 4　APP 首页与“课程表”服务

Fig. 4　UI of the APP and the curriculum schedule service



（a）识别样例　　（b）准确率统计

（a）Examples　　（b）Statistics

图 5　验证码识别服务测试效果

Fig. 5　Test for the identification of verification code

4.3　应用与推广

为进一步说明本文所述的解决方案及相应系统的实用性，通过微信小程序数据服务获得了 2018 年 9 月至 12 月，APP 用户和访问数统计详情，如图 6 所示。可见，本系统用户数稳定上升，访问次数以 7 天为单位呈周期地波动，完全符合高校教学周的活动特征。

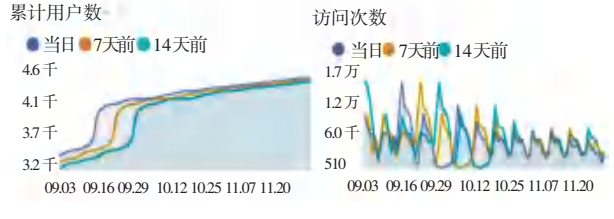


图 6　本系统应用情况统计（提取自“微信小程序数据助手”）

Fig. 6　Statistics of the application of the implemented system (supplied by the WeChat applet data assistant)

（下转第 124 页）