“互联网+”环境下移动校园搜索引擎设计与实现

王宁邦,徐 博
(云南师范大学 信息管理处,云南 昆明 650500)

摘 要:“互联网+”环境下重新思考高校移动网络文化建设、整合门户信息及其传播问题显得很有必要。综述了关于高校信息门户整合以及移动校园网络文化建设现状。提出“互联网+”环境下高校移动校园搜索引擎设计并对原型系统进行实现。高校移动校园搜索引擎内容包括:信息服务物理模型、整合数据源提供一站式信息服务、主题爬虫技术、累积网络行为,维系大学校园和学生情谊、挖掘平台隐性业务促进网络文化育人、“互联网+”环境下高校移动校园搜索引擎系统特点。对前端实现过程及技术、WebKit的渲染过程、规则爬虫数据采集等关键技术进行介绍,原型系统的网站采集、Android端和微信端显示效果进行呈现。通过高校移动校园搜索引擎建设对加强移动校园网络文化建设具有重要意义。

关键词:移动校园文化;校园搜索引擎;互联网+;主题爬虫;网站采集
中图分类号:311.5 **文献标识码**:A **文章编号**:1673-629X(2020)08-0157-07
doi:10.3969/j.issn.1673-629X.2020.08.027

Design and Implementation of Mobile Campus Search Engines in “Internet+” Environment

WANG Ning-bang, XU Bo
(Department of Information Management, Yunnan Normal University, Kunming 650500, China)

Abstract: It is necessary to rethink the construction of mobile Internet culture in universities and integrating and dissemination of portal information under the environment of “Internet+”. We summarize current situation of the integration of university information portal and construction of mobile campus network culture and propose the design and implementation of mobile campus search engines in “Internet+” environment. Content of mobile campus search engines includes information service physical model, integrated data source for one-stop information service, theme crawler technology, cumulative network behavior, maintaining university campus and student friendship, mining hidden business platform, promoting network culture education and specialty of mobile campus search engines system in “Internet+” environment. The front-end implementation process and technology, WebKit rendering process, rule crawler data acquisition and other key technologies are introduced. The website acquisition of the prototype system, Android and WeChat display effects are presented. It is of great significance to strengthen the construction of mobile campus network culture through the construction of mobile campus search engines in colleges and universities.

Key words: mobile campus culture; campus search engines; Internet+; theme crawler; website acquisition

1 概 述

“互联网+”代表着一种新的经济形态。“互联网+”是指以互联网为主的新一代信息技术,包括移动互联网、云计算、物联网、大数据等在经济、社会生活中各部门的扩散、应用与深度融合的过程[1]。2015年成为大数据发展的里程碑,在政府工作报告中,提出要制定“互联网+”计划,推动云计算、大数据与现代制造业的结合,促进大数据的升级发展。门户(portal)一词原意是指正门、入口,现多用于互联网的门户网站和企业应用系统的门户系统[1]。高校网站(Website)分为门户网站、二级院系或部门网站和专题网站[2]。文中信息门户为学校官网、学院或部门的主站。

移动无线互联网的时代已经到来,移动无线终端的数量已经超过有线终端,移动互联的应用需求日益

收稿日期:2019-08-03 **修回日期**:2019-12-06
基金项目:云南师范大学党建、思想政治工作理论研究项目(2018SZ22);云南师范大学2018年教师教育改革研究项目(JSJY201825);云南省教育科学研究基金项目(2019J0086);云南教育科学规划(高等学校教师教育联盟)教师教育专项课题(GJZ1905);云南省哲学社会科学教育科学规划项目(AFSZ19012)
作者简介:王宁邦(1985-),男,硕士,助理研究员,通信作者,主要从事信息安全与网络计算、计算机网络应用、软件工程、智能信息处理、电化教育、思想政治教育方面的研究。

增大,随着无线网络建设的发展和数字化校园应用系统的持续建设,校园移动终端应用已经逐步形成校园信息化的应用趋势。校园网各种应用向移动终端的迁移,提供真正适用移动校园网的应用服务,是每个学校面临的新挑战。显然,“一云多终端”风靡全网,单一的服务模式解决不了“互联网+”环境下的网络文化发展需求,移动网络文化融合主流应用如微信、主流移动技术、校园信息资源的发展势不可挡。QQ 空间文化、微信文化发展迅猛,借鉴它们的运作模式加强移动校园网络文化建设具有一定的研究意义。

当下,移动校园网络文化明显出现建设缺位、没有吸引力、隐形外流等情况,信息门户也往往回避不了无人问津的尴尬,融入主流技术整合高校信息门户资源,丰富、挖掘校园网络文化的承载渠道,加强高校移动网络文化建设,研究门户信息网络传播规律以及网络舆情分析与引导能力显得格外重要,在“互联网+”环境下重新思考高校移动网络文化建设、整合门户信息及其传播问题显得很有必要。创新符合网络传播规律的网上宣传方式,提升网络舆情分析和引导能力。加强互联网分类管理,强化运营主体的社会责任。推进文明办网、文明上网,引导广大青年争当“中国好网民”,倡导网络公益活动,净化网络环境。可见高校信息门户将会被融入时代的主流技术。同样的,高校移动校园网络文化也需要依托移动端移动技术、丰富的校园网络行为等的承载[3]。

“分久必合”,各学院门户自成一家,信息服务不集中,快捷查询门户信息显得不方便。就桌面而言,缺少一键查询获取所需信息的应用,虽然百度可以做到这一点,但是由于它的工作量很大,无暇顾及校园门户信息;另外,桌面门户由于分辨率的问题,在移动端的门户信息显得不容乐观,但是移动端明显表现出来比桌面门户端更触手可及的优势。门户信息在移动端传播具有便利性,借助移动端 IOS、Android 等技术,个性化推送技术,第三方如微信等为桌面信息门户的发布提供便利,学生可以方便快速获取校园信息。网页自适应技术、HTML5 技术等可以让信息门户拥有更好的主流技术体验,然而现有的校园信息门户不具备这些体验[4]。同时,为每一个门户开发具备不错体验的客户端存在重复建设的问题,显得不太现实。所以融入主流技术统一提供门户信息服务、整合高校信息门户资源,丰富、挖掘校园网络文化的承载渠道,加强高校移动网络文化建设,研究门户信息网络传播规律以及网络舆情分析与引导能力显得格外重要。

“互联网+”环境下高校移动校园搜索引擎相关研究现状分析如下:(1)信息门户整合方面;唐宏平[1]认为信息门户具有“统一管理信息资源、信息技术整合和信息共享”的优势,并研究与应用信息门户技术等搭建起了辽河油田新的集中统一的信息门户系统。马国良[2]基于 Web 服务及其关键技术(HTTP、XML、SOAP 等)、门户技术等建立统一门户。方玲慧[5]针对目前美国高校门户网站建设的现状和存在的问题,对加强网站建设的对策和方法进行探讨。方伟杰[6]通过在数据整合与身份认证整合的基础上进行信息整合实现高校资源与服务的综合利用。周晓艳[7]将一卡通系统纳入信息门户平台中,丰富信息平台内容。付小龙[8]阐述了信息构建理论在数字校园信息门户规划与设计的指导作用。林丽娟[9]提出信息整合的分层整合架构。蓝鹰[10]提出了一套基于 HTML5+CSS 技术的高校门户网站生成方案。毕剑[11]采用响应式网页设计技术,为图书馆移动门户的建设提供了一种新的解决方案。关于信息门户整合的文章较少,百度、Google 等技术明显产生了很好的经济效益和社会效益,随着移动技术的发展,考虑“互联网+”环境下,高校利用搜索引擎等技术对信息门户整合具有重要研究意义。(2)高校移动网络文化建设方面;孙耀庭[12]对开放大学移动校园 APP 服务功能需求进行探索。燕玲玲[13]基于 Android 平台,建立一个针对本校学生的实时校园生活信息服务系统,提供学生交流互动的平台,打破传统的只能上贴吧论坛交流的格局。(3)网络爬虫技术方面;岳雨俭[14]提出基于 Hadoop 分布式网络爬虫技术,具有较高的抓取效率。(4)基于用户偏好的个性化推送方面;黄原原[15]提出一个基于百度社区和领域本体库,结合相关反馈技术和扩展查询技术,促使个体特征库不断学习用户知识以提供个性化信息检索的模型。周蒙[3]利用信息推送技术、个性化广告推送技术等,设计并实现个性化广告推送服务系统。刘思源[4]设计并实现了一种基于用户偏好和地理位置信息的即时推送,并构建完整的个性化推送模型。

综上所述,当前的信息门户资源服务方式存在效率低下的问题,尤其是大数据释放红利的时代,其次,分散建设集中服务显得具有更大的影响力,产生了更大的社会效益,所以利用前沿的计算机技术、网络技术、云计算技术整合校园网门户资源,优化信息门户服务效率的移动校园搜索引擎相关研究具有重要意义。在“互联网+”环境下重新思考高校移动网络文化建设、整合门户信息及其传播问题显得很有必要。此外,校园网络文化平台构件缺失,文化资源外流严重,由本校学生形成的纯净校园网络文化氛围亟待形成,移动校园网络文化平台承载学生和学校信息门户的发展,可以依托移动校园网络文化个人空间,记录学生在学校的网络文化行为,充分挖掘高校信息资源与学生网络行为之间的关系,为构建“互联网+”环境下移动校

园网络文化平台奠定基础。

2　“互联网+”环境下高校移动校园搜索引擎设计

以云南师范大学各个学院信息门户为实例对象,拟研究整合校园门户信息,提供方便快捷、融合移动体验、网页自适应技术、HTML5 技术的掌上信息服务,提供一键查询校园搜索引擎的终端功能,让师生体验到真正的移动校园。开通移动校园文化个人空间,进一步构建、繁荣移动校园网络文化。

2.1　信息服务物理模型

从桌面网页到移动端数据,涉及对原始桌面网页主要信息的获取,由网址获取页面所有内容,再通过设定的通用规则对所需要的数据进行抓取并存储,为了提供信息构建的数据访问接口以及客户端请求程序等,处理流程如图 1 所示。

2.2　“互联网+”环境下高校移动校园搜索引擎主要内容

高校移动校园搜索引擎系统结构如图 2 所示。

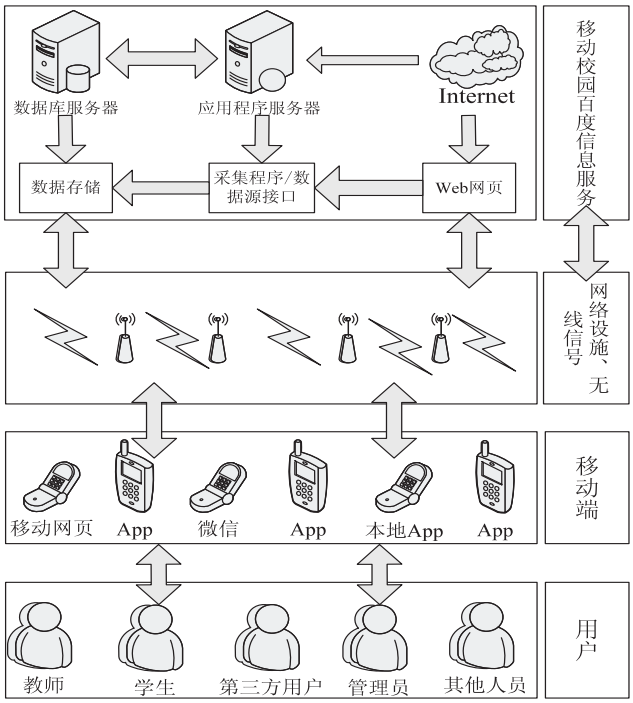

图 1　信息服务物理模型

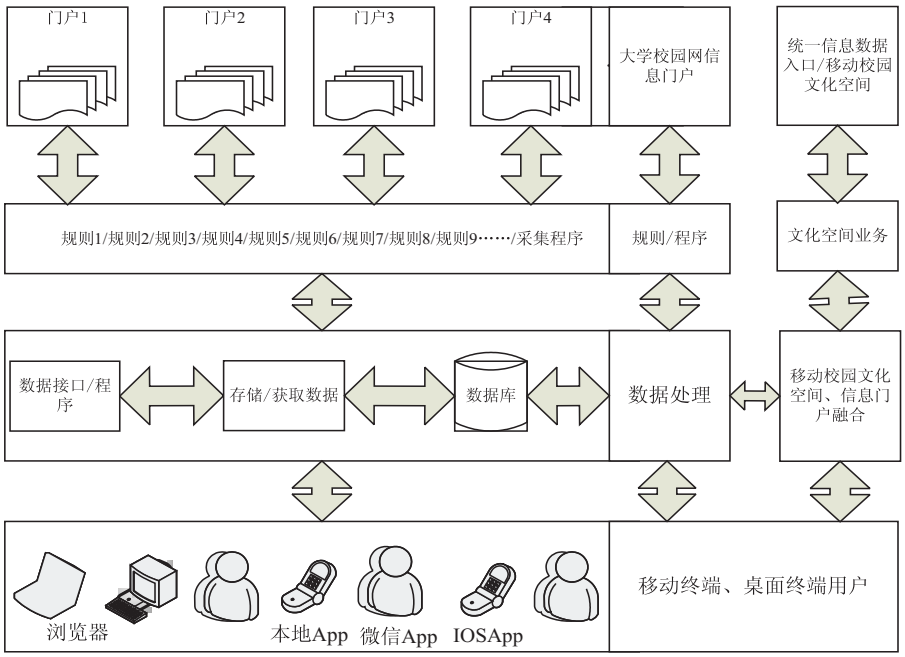

图 2　高校移动校园搜索引擎系统结构

2.2.1 整合数据源提供一站式信息服务

以云南师范大学为例,在数据源方面,让每一个门户提供数据访问接口显得繁琐,研究依托学校的云计算平台,利用爬虫定时对 100 多个校园门户信息进行爬取,构建校园搜索引擎索引库。针对就业处,拟研究爬虫在互联网爬取就业信息,为毕业生提供完备的就业信息源。

提供网页自适应门户:研究兼容桌面和移动端的网页自适应技术以及 HTML5 技术,为桌面、第三方如微信等提供数据接口,搭建“一键查询所需”的统一入口。使用校园搜索引擎的移动客户端,为师生提供一键移动校园、触手可及的门户信息服务。研究设定数据采集规则,裁剪信息门户冗余成分,增加移动端技术元素,让门户信息服务拥有移动体验。

2.2.2 主题爬虫技术

图 3 中,黑色节点为主题相关网页,白色节点为主题无关网页,Community Q 为许多主题无关网页组成的区域。假设爬虫从 P0 开始爬行,理想的主题爬虫,应该能够预测网页的主题相关性,沿着图中箭头所指的方向爬行,剪掉不相关网页,舍弃 P3 这个分支,尽可能少地下载不相关网页;并且准确判断出 P2、d0 等的主题相关性,抓取到这些网页。主题爬虫工作流程见图 4。

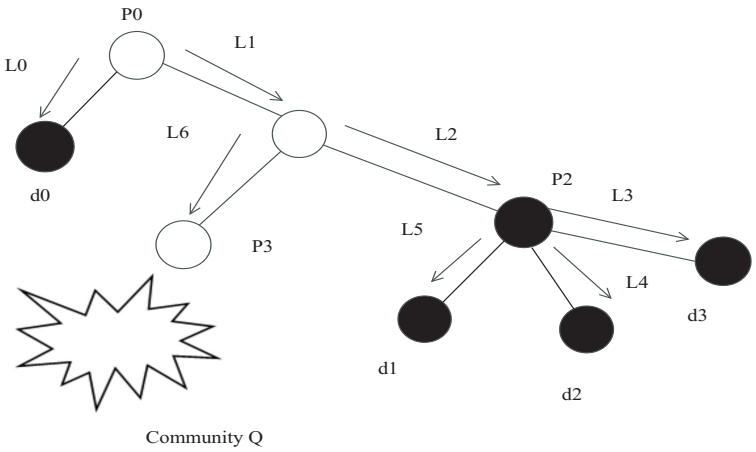

图 3 网络爬虫搜索示意图

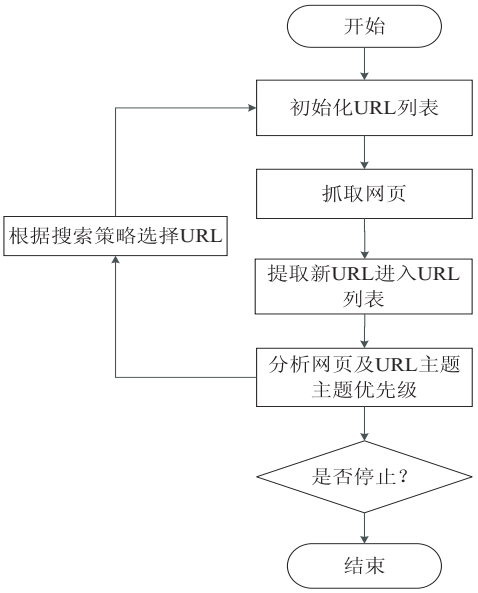

图 4 主题爬虫工作流程

2.2.3 累积网络行为,维系大学校园和学生情谊

为学生提供移动校园网络文化空间,开通即时通信服务,记录网络行:如即时通信信息、校园空间心情等,为加强移动校园网络文化建设累积基础数据,统计“校园最文化”:最受关注的部门、最后关注的学生、最受关注的话题等,基于位置、轨迹等研究“我与我的校园”,统计毕业生对学校的关注度等。

2.2.4 挖掘平台隐性业务促进网络文化育人

在逐步完成的基础上上线运行测试,对用户行为进行挖掘形成新的业务。并且研究网络传播规律的网上宣传方式,提升网络舆情分析和引导能力,推进文明办网、文明上网,引导广大青年争当“中国好网民”,倡导网络公益活动,净化网络环境,在原型开发的过程中逐步将这些需求一一落实到每一个功能业务。如对用户网络行为(如对校园信息门户建设的点击贡献率)按照一定的标准量化核算成分数,给以“师大好网民”的电子奖励,给予团学积分奖励等。

3 “互联网+”环境下高校移动校园搜索引擎系统特点

以“互联网+”为研究背景,使用自适应网页、HTML5、移动端等主流技术整合高校信息门户,构建高校移动网络文化平台,提供一键移动搜索引擎校园、一掌移动校园文化体验。其次,项目将充分利用超链接技术、关键词技术来设计开发移动校园搜索引擎。网络育人、维系学生终身与大学校园文化这条纽带、最关心本科生等情感目标的移动校园文化,使用计算机

技术实现需要过程的定义,尤其是在师范院校,如预测校园突发事件的发生、识别等。

(1)顺应政策指向;充分以“互联网+”为大背景,项目依照“十三五”规划的指示,将计算机科学领域主流技术应用到高校信息门户整合提供移动校园信息门户服务,加强移动校园文化平台建设,融合信息门户整合以及移动校园文化移动建设,兼顾了加强现代传媒体系以及加强网络文化建设的目标,移动校园个人文化空间承载着学生在校园网络行为,如心情,符合云南师范大学建设“最关心本科生”大学的办校理念,同时也是大数据时代的到来要求每一个领域均掌握第一手大数据的迫切需求。

(2)“一键搜索引擎校园,一云多终端”体验;整合校园信息门户数据,依托计算机领域主流技术提供一站式搜索引擎校园服务。一云多终端校园搜索引擎让校园信息门户无处不在、触手可及,校园移动文化空间和校园门户的信息传播相辅相成,让信息门户和校园移动文化空间深度融合、相互共生。

(3)丰富的校园网络资源;校园网络有丰富的硬件资源、软件资源,尤其是项目可以依靠云南师范大学的云计算平台,可以利用这样的软硬件环境实现高性能计算以及处理高并发量。同时,由于各种服务器资源均在校园网内,校园内网不需要依赖 Internet,保证了高速的网络体验。

(4)集群门户的移动校园搜索引擎信息服务模型;模型结合了主题爬虫技术、个性化推荐算法等对校园网络信息门户进行加工,为校园用户提供个性化的信息服务。

4　关键技术

移动校园搜索引擎系统技术路线如图 5 所示。

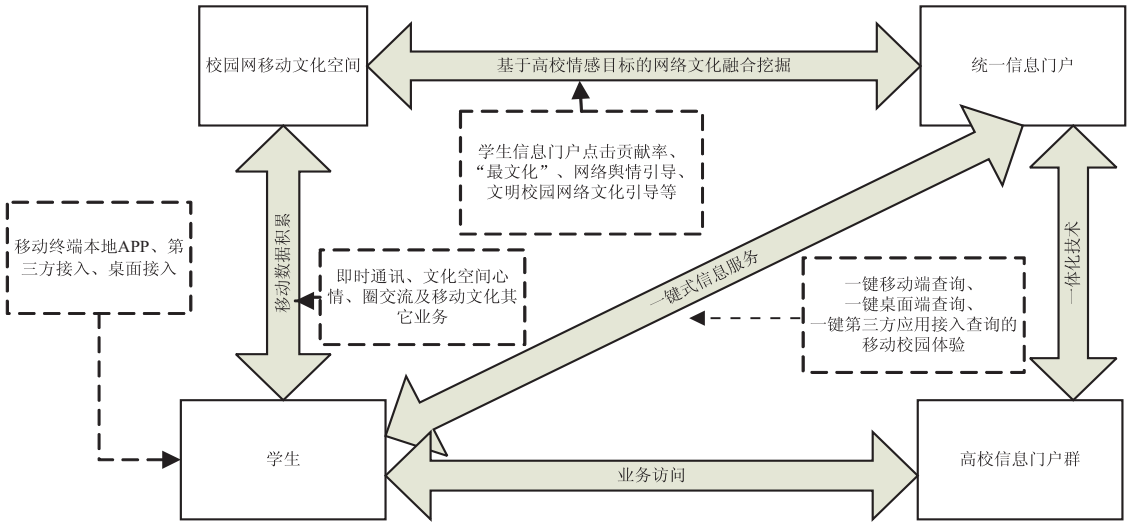

图 5　移动校园搜索引擎技术路线

系统涉及学生、统一信息门户、高校信息门户群、校园网络文化移动空间。系统是从零平台到构建平台,从无数据源到构建数据源,从抽象的“最关心本科生”、移动校园网络文化、舆情引导等抽象概念到业务、逻辑、技术、融合的实现,从基础统计到深度算法挖掘,从门户和文化空间的无关联到相互映衬,从线下文化到线上文化的隐形挖掘再到校园网络文化繁荣的过程。

4.1　前端实现过程及技术

文中主要应用了移动端 Webkit、服务端数据采集存储、计算机网络等技术。移动的应用程序如浏览器、个性化应用主要依靠 Webview,Webview 的引擎是 Webkit,Webkit 是苹果发起的一个开源项目,还对 HTML5 提供支持。

4.2　Webkit 的渲染过程

HTML 在移动端表示的过程是 Webkit 的网页渲染过程,第一阶段是从 URL 到构建完 DOM 树,第二阶段是 DOM 树到构建完 Webkit 的绘图上下文,第三个阶段是从绘图上下文到最终生成的图像,具体为:将网页内容,送到 HTML 解释器。HTML 解释器在解释它后形成 DOM 树,中间如果遇到 JavaScript 代码则交给 JavaScript 引擎去处理。如果页面包含 CSS,则交给 CSS 解释器去解析。当 DOM 建立的时候,接受来自 CSS 解释的样式信息,构建一个新的内部绘图模型。该模型由布局模块计算模型内部的各个元素的位置和大小信息,最后由绘图模块完成从该模型到图像的绘制。

在网页内容的下载中,需要使用到网络和存储。计算布局和绘图的时候,需要使用 2D/3D 的图形模块,同时因为要生成最后的可视化结果,这时候需要开始解码音频、视频和图片,同其他内容一起绘制到最后的图像中。

4.3 规则爬虫数据采集

(1)设定规则:现有桌面网页均是很有规律的页面,而且除了具有动态信息外相对固定,根据现有桌面网页特点设置网页采集规则,如使用正则表达式获取各种标签,超链接、标题等,往往呈现大类下面具有小类,小类下面才有标题,标题通过超链接得到网址,每一个阶段都可能需要正则表达式等的支持,才能获取到所需要的内容;

(2)采集数据:根据设定好的规则采集满足规则的数据;

(3)产生数据源:将采集到的数据进行存储以便移动端数据请求需要,或者为了节省数据库资源,不再对数据进行数据库存储,和对数据进行存储相比较,直接提供数据源可以根据每一次移动端的需求启动采集程序获取相应数据;

(4)移动端请求数据:移动端根据用户具体点击的模块,向数据源获取所需要的数据,并对数据使用相应空间给以呈现;

(5)自适应页面布局,在保证能够抓取到信息详情关键内容及其 HTML 标签的同时,由于捕获的内容中如图片是适合 PC 端呈现的,但是在移动端还是不能自适应,在采集到的数据里面添加自适应标识,如使用 CSS 进行图片的自适应代码。

5 移动校园搜索引擎原型系统

随着终端技术的不断发展,通用自适应信息门户采集及展示系统可以用于现有的、不具有自适应功能的信息门户等,为“一云多终端”信息展示提供解决思路。根据门户信息等的特点,设定采集规则,存储具有自适应多终端元素的门户信息,提供数据源接口、微信和本地 App 等的展示,移动校园搜索引擎原型系统以云南师范大学信息管理处门户网站为例。移动校园搜索引擎原型系统功能包括:

(1)云南师范大学信息管理处门户信息自适应采集:根据门户网页设定规则采集自适应的门户信息;

(2)云南师范大学信息管理处门户信息存储:设计数据库结构对采集到的门户自适应信息进行存储;

(3)云南师范大学信息管理处门户信息展示:①基于 Android 的本地 App 门户展示;②基于微信的门户信息展示。

移动校园搜索引擎原型系统技术特点:

(1)采用混合开发模式,其中包括基于 Android 本地应用、微信等的移动端展示,以及数据采集的 C/S 数据源服务端;

(2)具有“一云多终端”特点。基于服务端采集的一个自适应数据源可以为主流平台 Android、微信等提供数据准备,避免了普通网页在移动端呈现混乱的情形;

(3)通用性。系统具有普适性,其他具体应用可以更改采集规则、数据表结构等进行套用。

编程语言及其版本号:Java7、Android 4.0、Mysql 5.6.24、Php 5.3.29。

5.1 网站采集界面

界面在运行过程中,以 Loading 作为提示,同时,日志窗口抓取运行产生的结果,客户端启动后,每隔设定好的时间间隔重复运行过程。采集运行界面如图 6 所示,下一次运行开始时的采集完成或间隔界面如图 7 所示。

图 6 采集运行界面

图 7 采集完成或间隔界面

5.2　Android 端、微信端效果图

基于 Android 的移动端呈现网站导航栏目标题信息主界面、基于 Android 的移动端对某条消息详情呈现界面效果良好。微信端访问接口界面、门户主目录、详情界面、目录下内容标题列表界面如图 8 和图 9 所示。

图 8　门户目录

图 9　目录下内容标题列表

6　结束语

系统采用自适应网页设计、HTML5、移动端(Android、IOS 等)、第三方如微信等主流应用或技术,以云南师范大学校园网络信息门户为对象,使用自行设计的规则爬虫,基于高性能、多并发的云计算平台提供应用服务和存储服务支持,获取到的数据为挖掘构建移动校园文化的计算机实现提供保障,而且项目系统模型构建与设计在前期工作中已经通过几个门户测试证明可行。同时,移动校园文化与现有信息平台高度融合并服务于舆情监测与控制、网络行为预测具有相关理论支撑。做好上线运维工作,并在此过程中继续挖掘构建移动校园网络文化的业务,对平台进行网络推广,为进一步丰富移动校园文化奠定基础。研读个性化推荐算法,获取适合移动校园搜索引擎个性化信息服务,利用实验法选择预先设定的主题爬取门户资源,实现个性化相关模型如用户模型等,挖掘用户偏好,接受统一资源的集中个性化服务。实际对接校园网络门户信息资源数据,实测原型系统的信息资源爬取以及个性化信息服务是下一步的研究方向。

参考文献:

[1] 唐宏平. 信息门户迁移整合系统的设计与实现[D]. 成都:电子科技大学,2010.
[2] 马国良. 基于 Web 服务的信息系统集成研究与应用[D]. 长春:吉林大学,2013.
[3] 周　蒙. 面向互联网用户的个性化广告推送服务研究——基于 Hadoop[D]. 上海:东华大学,2014.
[4] 刘思源. 基于 Android 的信息分享系统及个性化推送的设计与实现[D]. 北京:北京邮电大学,2013.
[5] 方玲慧,林安琪. 美国高校门户网站的现状分析[J]. 出国与就业,2012(6):185.
[6] 方伟杰,洪　波,王建国,等. 信息整合在高校信息门户建设中的应用[J]. 福建电脑,2011,27(8):141-142.
[7] 周晓艳. 校园信息门户中一卡通数据的集成[J]. 科技信息,2010(26):388.
[8] 付小龙,肖　平,袁　芳,等. 数字校园信息门户信息构建方法的研究与实现[J]. 中国教育信息化,2010(11):12-15.
[9] 林丽娟. 数据中心在学校信息整合中的应用研究[J]. 新余高专学报,2010,15(2):87-89.
[10] 蓝　鹰. 基于 HTML5 技术的高校门户网站设计[J]. 智能计算机与应用,2015,5(4):95-96.
[11] 毕　剑,刘晓艳,张　禹. 使用响应式网页设计构建图书馆移动门户网站——以云南大学图书馆为例[J]. 现代图书情报技术,2015(2):97-102.
[12] 孙耀庭,陈　信. 开放大学“移动校园”构建的探索[J]. 中国教育信息化,2007(10):7-9.
[13] 燕玲玲. 基于 Android 的高校校园通的设计与实现[D]. 太原:山西大学,2013.
[14] 岳雨俭. 基于 Hadoop 分布式网络爬虫技术的研究[D]. 淮南:安徽理工大学,2015.
[15] 黄原原. 百度搜索技术及其个性化信息搜索探析[J]. 农业图书情报学刊,2010,22(2):84-87.