基于网络爬虫的文献检索系统的研究和实现

杨 洋[1,2],李晓风[1,2],赵 赫[1,3],刘 冰[1,2]
(1. 中国科学院 合肥物质科学研究院,安徽 合肥 230031;
2. 中国科学院大学,北京 100049;
3. 中国科学技术大学,安徽 合肥 230026)

摘 要:文中系统基于网络爬虫技术实现了文献资源的智能搜索和关键信息的抓取功能,把采集到的信息采用本体论的方法进行分类识别,并自动存储文献资源到本地服务器。下载子系统采用负载均衡的方法把下载任务分配到多个服务器。系统采用高效的 Protobuf socket 通信手段,提供高效准确的内部下载服务。通过对内提供统一门户入口的方式对检索和下载行为进行记录,有效避免了同一资源的重复下载,也使得文献检索和下载行为变得可追溯,为图书文献情报管理和研究工作提供了数据支撑。该系统可有效减少科研机构获取学术资源所需的资金投入并减少网络带宽占用。

关键词:网络爬虫;本体论;论文检索;Web;MVC;负载均衡
中图分类号:TP393.4 **文献标识码**:A **文章编号**:1673-629X(2014)11-0035-04
doi:10.3969/j.issn.1673-629X.2014.11.009

Research and Realization of Academic Search System Based on Network Crawler

YANG Yang[1,2],LI Xiao-feng[1,2],ZHAO He[1,3],LIU Bing[1,2]
(1. Hefei Institutes of Physical Science,Chinese Academy of Sciences,Hefei 230031,China;
2. University of Chinese Academy of Sciences,Beijing 100049,China;
3. University of Science and Technology of China,Hefei 230026,China)

Abstract:This system has realized intelligent search and external academic resources capture based on network crawler technique. It uses ontology technology to identify each article and automatically store the resources into local repository. Downloading subsystem in this system applies load balance method to distribute downloading tasks equally to each download server. Protobuf,a high-efficiency communication mechanism,provides downloading service with high availability and accuracy in this system. At the same time,this system has solved the problem of repeated downloading and access recording by offering a unique entrance to the whole institute. Access control is also designed to eliminate malicious and excessive downloading. System automatically saves user searching data,which makes information retrieval becomes traceable,providing data support for library information management and research. This system can effectively reduce expense on digital academic resources for institute and network bandwidth.

Key words:network crawler;ontology;thesis retrieval;Web;MVC;load balancing

0 引 言

为了获得所需的学术资源,科研机构每年都需支付大量费用给电子学术资源服务提供商。例如作者所在单位从 2010 到 2013 短短 3 年间电子学术资源费用增长了 1.914 倍。由于电子学术资源一般是根据下载量和阅读量计费,所以重复购买相同的学术资源造成了资金的很大浪费。怎样避免学术资源的重复下载是科研机构十分关注的问题。

针对这一问题,本系统通过对学术资源提供商的网站研究和分析,实现了智能搜索和文献资源下载[1]。由于下载服务器具有网络带宽优势,并且部分论文已下载到本地服务器,下载速度较之前得到明显提高。系统的应用可帮助科研机构减少为获取学术资源所需的资金投入,也可有效减少网络带宽占用。

收稿日期:2013-12-30 **修回日期**:2014-04-07 **网络出版时间**:2014-09-11
基金项目:中国科学院重点项目(院 1221)
作者简介:杨 洋(1990-),男,江西九江人,硕士研究生,研究方向为软件工程;李晓风,博士生导师,研究方向为计算机应用和网络安全等。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140911.1009.042.html

1 系统设计

1.1 系统架构

系统包含了两个子系统,即 Web 服务系统和论文下载系统,分别部署于不同的服务器以减小服务器的压力。系统组成架构如图 1 所示。Web 服务系统基于. NET MVC 提供 Web 服务,实现信息记录、关键字的搜索、关键信息的抓取、论文一致性检测等功能。下载子系统基于 Java 和 Protobuf socket[2],提供高速的论文下载功能。

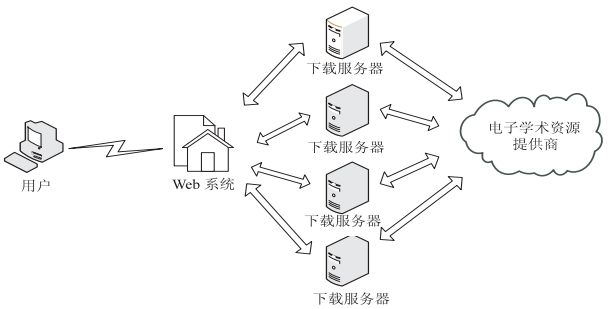

图 1 系统架构图

Web 子系统采用了. NET MVC 框架开发,. NET 框架是微软的统一技术平台,开发人员用不同的语言开发的程序被编译成微软中间语言后可以在任何微软的平台上运行,提高了开发效率和代码的复用性。而 MVC 是一种在图形化界面程序中很流行的架构设计模式,MVC 是 Model(模型)、View(视图)及 Controller(控制器)的缩写。正因为 MVC 在其他语言获得了巨大的成功,微软也响应. Net 开发人员的期待推出了. NET 的 MVC 框架,使用. NET 的 MVC 框架进行 Web 开发时能高效地实现逻辑和前端展现的解耦,使得前端开发和后台逻辑能很好地隔离,降低了程序开发和后期维护的成本。

Java 是一种可以撰写跨平台应用软件的面向对象的程序设计语言。Java 技术具有突出的通用性、高效性、平台移植性和安全性,同时拥有全球最大的开发者社区。为了后期能够部署在不同的平台上构成一个异构的分布式平台,下载服务器选择了 Java 进行开发。

1.2 系统流程图

系统流程图如图 2 所示。

1.3 算法的分析

(1)网络爬虫。

网络爬虫是一个抓取网页内容的程序,利用网页格式特征进行网页分析[3]。系统利用网页的标签结构分析出论文的相应信息,如标题、摘要、关键字等。为了提高抓取效率和准确度,系统内的网络爬虫有针对性地做了一些优化改进[4]。如一些热门关键字往往会被反复检索,就没有必要每次都重复爬取搜索结果,因此系统在服务器端这些热门搜索结果进行缓存处理,提高了系统运行效率。爬虫抓取的内容依赖于网页格式,为了将爬虫行为与网页格式解耦,系统将网页格式信息抽象为配置文件,在运行时读取配置来定制爬虫的行为,使得系统可以适应网页格式的变化[5]。

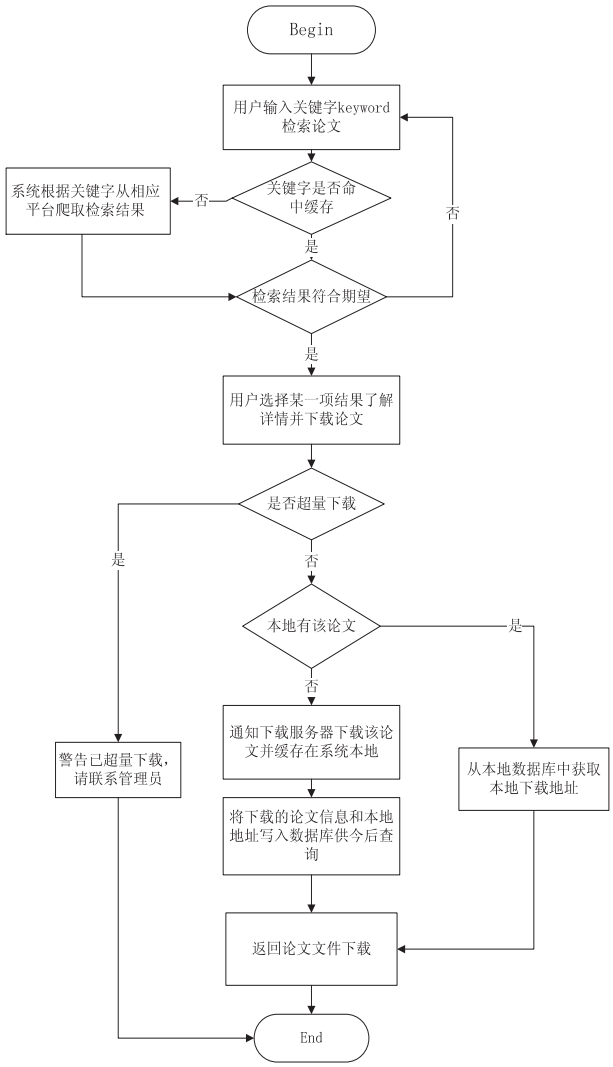

图 2 系统流程图

(2)论文唯一性识别。

系统必须能唯一识别一篇论文,才能判断用户提出下载的某篇论文是否已缓存在本地。该系统基于本体论[6-8]的方法,对论文对象进行抽象,建立了论文的本体模型,声明了论文的元数据结构,如表 1 所示。

表 1 论文的本体模型

元素	字段名	类型	说明
标题	title	字符串	论文的标题
作者	author	字符串	论文的作者
摘要	summary	字符串	论文的摘要
关键字	keywords	字符串	论文的关键字
出版时间	publish_time	日期	论文的出版时间
出版期刊	press	字符串	论文的出版期刊

系统逐一比较每个元素,判断用户提出下载的论文与本地某篇论文是否为同一篇论文,如果比较结果

完全相同,则系统判断为同一篇论文,返回本地的下载地址。如果比较结果不相同,系统就通知下载服务器下载该论文到本地服务器[9-11]。

2　系统功能的实现

2.1　下载身份验证

部分电子学术资源服务商在通过 IP 地址进行授权访问的同时,还要求在 IP 地址范围内的用户提供用户名密码。因此下载服务器在请求下载页面时需要将用户名密码通过 POST 方式发送给下载页面,验证成功后,电子学术资源服务商才会提供论文下载。以下是具体实现的 Java 代码:

```
URL url = new URL(fileUrl);
HttpURLConnection con = (HttpURLConnection) url.openConnection();
con.setRequestMethod("POST");
String urlParameters = "username = * * * &password = * * * ";
DataOutputStream wr = new DataOutputStream(con.getOutputStream());
wr.writeBytes(urlParameters);
wr.flush();
```

2.2　搜索结果和论文关键信息抓取

为了实时地搜索论文的关键信息,系统把用户输入的关键字发送到电子学术资源服务器处理,获取返回搜索的结果后解析成论文实体信息,显示到 Web 页面上展示给用户。下面是某个学术资源提供商某个检索结果页面的 html 代码。

```
<div class="wz_tab">
<div class="wz_content">
<h3>
<a href=[论文详情页面地址]>[标题]</a><a href=[论文下载地址]><img src="download-icon.jpg"/></a>
</h3>
<div class="width715">
<span class="text">[论文摘要]</span>
</div>
<span class="year-count">[论文发表年份]</span>
<span class="count">[论文下载次数]</span></span></div>
</div>
```

根据网页内容的分析算法以及不同的标签,分析出论文的相应信息,如标题、摘要和下载次数。本系统利用 htmlagilitypack[12] 和 XPATH 语法[13] 将 html 页面成功地解析为论文类实体,通过友好的页面展现给用户,效果如图 3 所示。

当用户对某篇论文进一步查看详情时,系统再次根据网页内容的分析算法和网页的标签格式,分析出论文更加详细的信息。如论文摘要、论文关键字、论文分类号、论文目录等。图 4 是解析详细页面后的运行效果。

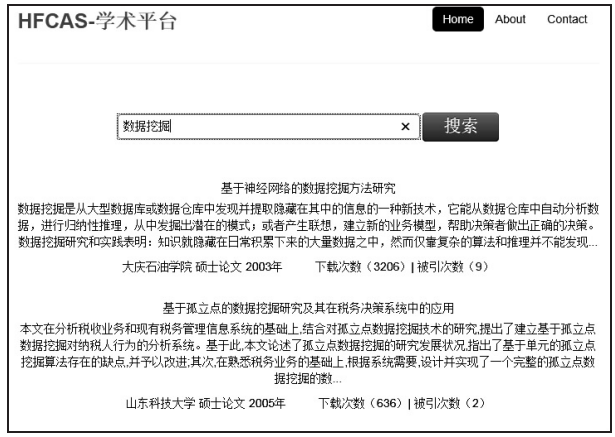

图 3　搜索结果页面

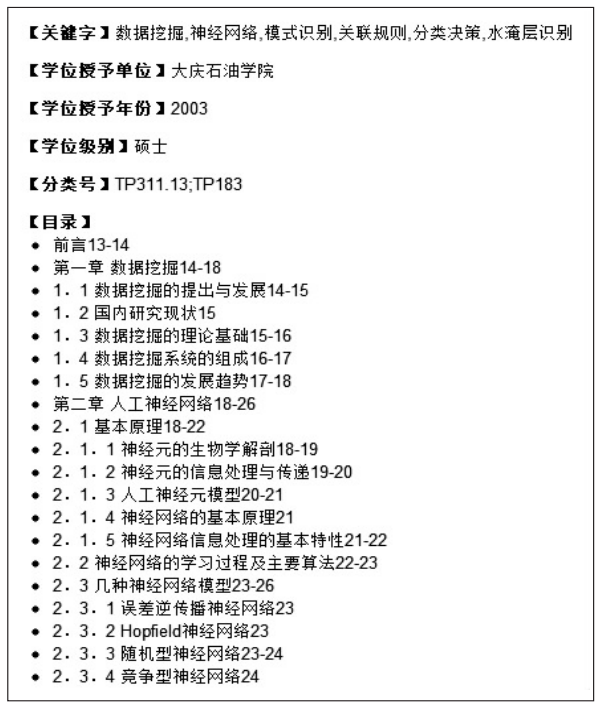

图 4　论文详情页面

2.3　论文的下载

当用户点击下载某篇论文时,Web 服务系统会基于论文的本体唯一地检索识别论文,判断出论文是否存在于本地的服务器,如果存在则直接从本地下载;如果本地服务器中不存在该文献,则通知下载服务器进行实时下载。当下载服务器下载完成后,将文献文件传送给用户。

2.4　记录学术资源数据

系统使用 MySQL 数据库记录每次检索和下载的相关信息,如检索关键字、检索源 IP 地址、下载论文实体信息、下载次数等。记录下每个用户的下载数据后就可以提前预防恶意过量下载。同时为后期的图书文献情报管理和研究工作提供数据依据。

3 结束语

文中系统在内部某中心试运行一个月以后,共接收了 3 125 次下载请求,这其中包含了 871 次重复下载,系统实际下载 2 254 篇论文。节省了约 27.87% 的下载量。另外系统在使用过程中积累的访问数据和下载数据后期可以用来做数据挖掘,用于分析机构内学术趋势和学术状况,这对于科研机构的长期发展也有一定的实际意义[14-15]。

系统暂时只覆盖了部分学术资源库,后期需要增加学术资源的覆盖面。系统中保存了很多有价值的使用数据,如何有效地利用这些数据挖掘出有价值的信息将是这个系统在推广使用过程中下一步值得思考和解决的问题。

参考文献:

[1] 李育嫦. 文献检索中提高查全率与查准率的方法探讨[J]. 图书馆学研究,2002(11):92-93.

[2] 李纪欣,王 康,周立发,等. Google Protobuf 在 Linux Socket 通讯中的应用[J]. 电脑开发与应用,2013,26(4):1-5.

[3] 詹恒飞,杨岳湘,方 宏. Nutch 分布式网络爬虫研究与优化[J]. 计算机科学与探索,2011,5(1):68-74.

[4] 尹 江,尹治本,黄 洪. 网络爬虫效率瓶颈的分析与解决方案[J]. 计算机应用,2008,28(5):1114-1116.

[5] Stevanovic D,An Aijun,Vlajic N. Feature evaluation for Web crawler detection with data mining techniques[J]. Expert Systems with Applications,2012,39(10):8707-8717.

[6] 朱庆生,邹景华. 基于本体论的论文检索[J]. 计算机科学,2005,32(5):172-173.

[7] 邓志鸿,唐世渭,张 铭,等. Ontology 研究综述[J]. 北京大学学报(自然科学版),2002,38(5):730-738.

[8] Abruscia V M, Fouqueré C, Romanoa M. Formal ontologies and coherent spaces[J]. Logic Categories Semantics, 2014(1):67-74.

[9] Yan Wei,Zanni-Merkb C,Cavalluccia D,et al. An ontology-based approach for inventive problem solving[J]. Engineering Applications of Artificial Intelligence,2014,27:175-190.

[10] van Ruijven L C. Ontology for systems engineering[J]. Procedia Computer Science,2013,16:383-392.

[11] Liu Chilun. Cloud service access control system based on ontologies[J]. Advances in Engineering Software,2014,69:26-36.

[12] Simonm. HtmlAgilityPack's documentation[EB/OL]. 2006. http://htmlagilitypack. codeplex. com/.

[13] 万维网联盟(W3C). XPATH 语法介绍[EB/OL]. 2010. http://www. w3school. com. cn/xpath/.

[14] 李 洁,丁 颖. 语义网、语义网格和语义网络[J]. 计算机与现代化,2007(7):38-41.

[15] 李 蕾,郭祥昊. 基于语义网络的概念检索研究与实现[J]. 情报学报,2000,19(5):525-531.

(上接第 34 页)

工作包括继续提高差异策略的精细化,模拟交互效果的真实性,数据收集的有效性等。

参考文献:

[1] 方志鹤. 恶意代码分类的研究与实现[D]. 长沙:国防科学技术大学,2011.

[2] Android malware genome project[EB/OL]. 2010. http://www. malgenomeproject. org/.

[3] Zhou Yajin,Jiang Xuxian. Dissecting Android malware:characterization and evolution[C]//Proc of 2012 IEEE symposium on security and privacy. San Francisco:IEEE Computer Society,2012:95-109.

[4] Zhou Yajin,Wang Zhi,Zhou Wu,et al. Hey,you,get off of my market:detecting malicious apps in official and alternative Android markets[C]//Proc of the 19th annual network and distributed system security symposium. [s. l.]:[s. n.],2012.

[5] Shin W,Kiyomoto S,Fukushima K,et al. A formal model to analyze the permission authorization and enforcement in the Android framework[C]//Proceedings of the 2010 IEEE second international conference on social computing. Minneapolis:IEEE,2010:944-951.

[6] Enck W,Ongtang M,McDaniel P D. On lightweight mobile phone application certification[C]//Proceedings of the ACM conference on computer and communications security. Chicago:ACM,2009.

[7] Barrera D,Kayaclk H G,van Oorschot P C,et al. A methodology for empirical analysis of permission-based security models and its application to Android[C]//Proceedings of the 17th ACM conference on computer and communications security. Chicago:ACM,2010.

[8] 张中文,雷灵光,王跃武. Android Permission 机制的实现与安全分析[C]//第 27 次全国计算机安全学术交流会论文集. 出版地不祥:出版者不详,2012.

[9] 丁丽萍. Android 操作系统的安全性分析[J]. 信息网络安全,2012(3):28-31.

[10] Shabtai A,Kanonov U,Elovici Y,et al. "Andromaly":a behavioral malware detection framework for android devices[J]. Journal of Intelligent Information Systems,2012,38(1):161-190.

[11] Burguera I,Zurutuza U,Nadjm-Tehrani S. Crowdroid:behavior-based malware detection system for Android[C]//Proc of 1st ACM workshop on security and privacy in smartphones and mobile devices. Chicago:ACM,2011:15-26.

[12] 马红素,郭燕慧. Android 应用自动化动态测试工具的研究及实现[D/OL]. 2012. http://www. paper. edu. cn.

[13] 王 伟. Android 病毒行为自动分析工具的设计与实现[D]. 天津:南开大学,2012.

基于网络爬虫的文献检索系统的研究和实现

作者： 杨洋， 李晓风， 赵赫， 刘冰， YANG Yang， LI Xiao-feng， ZHAO He， LIU Bing

作者单位： 杨洋,李晓风,刘冰,YANG Yang,LI Xiao-feng,LIU Bing(中国科学院 合肥物质科学研究院，安徽 合肥 230031；中国科学院大学，北京 100049)， 赵赫,ZHAO He(中国科学院 合肥物质科学研究院，安徽 合肥 230031；中国科学技术大学，安徽 合肥 230026)

刊名： 计算机技术与发展 ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2014(11)