**郑州大学毕业设计（论文）**

**题目：网络引文的爬虫系统初探**

**学生姓名 张磊（20133410139）**

**专 业 信息管理与信息系统**

**班 级 一班**

**指导教师 郭强（副教授）**

**学 院 信息管理学院**

**2017 年 5 月**

目录

[摘要 1](#_Toc482998989)

[Abstract 2](#_Toc482998990)

[1. 引言 3](#_Toc482998991)

[1.1. 选题的缘起和意义 3](#_Toc482998992)

[1.2. 研究内容和研究方法 3](#_Toc482998993)

[2. 相关概念解析 3](#_Toc482998994)

[2.1. 网络爬虫 3](#_Toc482998995)

[2.2. 运行环境 4](#_Toc482998996)

[2.3. 正则表达式 4](#_Toc482998997)

[3. 需求分析 5](#_Toc482998998)

[3.1. 业务需求描述 5](#_Toc482998999)

[3.2. 功能性需求分析 5](#_Toc482999000)

[4. 系统设计方案 6](#_Toc482999002)

[4.1. 系统结构 6](#_Toc482999003)

[4.2. 功能设计 6](#_Toc482999004)

[5. 系统实现 7](#_Toc482999005)

[5.1. 系统环境 7](#_Toc482999006)

[5.2. 网页下载与信息过滤功能 7](#_Toc482999007)

[5.3. 抽取有效数据 8](#_Toc482999008)

[5.4. 任务队列 9](#_Toc482999009)

[5.5. 数据的更新与存储 9](#_Toc482999010)

[6. 系统测试 9](#_Toc482999011)

[6.1. 抽取准确性 11](#_Toc482999012)

[6.2. 抽取用时 12](#_Toc482999013)

[7. 总结 12](#_Toc482999014)

[致谢 13](#_Toc482999015)

[参考文献 14](#_Toc482999016)

摘要

随着互联网技术的高速发展，网络信息资源越来越受到人们的重视，成为人们获取信息和进行科学研究的主要信息来源。网络引文，即网络参考文献也逐渐受到众多学术工作者的青睐，并成为一种不可忽视的引文类型。统计分析网络引文，不仅可以了解各学术领域对网络引文的引用现状，而且可以对其进行数据分析、数据挖掘。本文以万方数据库为例，通过爬虫抽取部分引用了网络引文的论文，为后续分析数据提供了强有力的数据保障。

**关键字**：网络引文 引文分析 数据挖掘 网络爬虫

Abstract

With the rapid development of Internet technology, more and more attention has been paid to network information resources. It has become the main information source for people to obtain information and carry out scientific research. Network citation, that is, network reference, has been gradually favored by many academic workers, and has become a kind of citation type which can not be ignored. Statistical analysis of network citation can not only understand the citation status of network citations in various academic fields, but also can analyze data and excavate data. This paper takes Wanfang database as an example, extracts some papers citing network citations through crawler, and provides a powerful data guarantee for subsequent analysis of data.

**Keywords**: Web Citations Citation Analysis Data Mining Web Crawler

1. 引言
   1. 选题的缘起和意义

近年来，计算机和信息技术有了突飞猛进的发展，产生了许多新概念和技术，如更高性能的计算机和操作系统、因特网（internet）、数据仓库（datawarehouse）、神经网络等等。在市场需求和技术基础都满足的情况下，数据挖掘的概念和技术应运而生。

数据挖掘（Data Mining）旨在从大量的、不完全的、模糊的随机数据中提取出人们事先不知道的潜在的有价值的信息或知识[1]。

网络的迅速发展使各行各业都开始采用计算机及相应的信息技术进行管理和运营。网络引文由于其传播方式的自由性、动态性，也逐渐成为论文参考文献的主要来源之一。为了能更方便的对论文中引用到的网络引文进行数据分析和数据挖掘，我们需要更方便的收集论文的网络引文，然而由于数据量较大、手工操作较为复杂，导致网络引文分析所需的数据供给不足，本文以此为出发点通过设计并实现一个网络引文爬虫系统来爬取数据，提供数据分析和数据挖掘所需的海量数据。

* 1. 研究内容和研究方法

网络爬虫系统主要是用来在浩瀚网络中收集网页或特定类型的数据信息，然后从中解析出客户所需要的信息，并持久化到数据库系统[2]。

本文以万方数据库为例，讲述网络引文爬取过程。首先介绍了网络爬虫的技术基础，然后研究了爬虫引文爬虫的一些关键问题如根URL的选取、URL消重等，最后实现了一个针对万方数据库的爬虫系统。

本文的主要研究内容如下：

1. 分析系统需求
2. 设计各功能模块
3. 描述爬虫系统的工作流程
4. 解析和保存数据
5. 测试数据
6. 相关概念解析
   1. 网络爬虫

网络爬虫（Network Spider），顾名思义是在网络中爬行的爬虫。其实质为一段计算机程序，该程序对数据的收集通过网页链接地址实现，往往从某一根URL出发，通过向远程服务器发起HTTP请求，读取网页内容，并从中获得与其相关的节点或子节点的链接地址，然后利用获取到的链接地址进行深入循环爬取[8]。

通用网络爬虫是从一个或几个初始URL开始，获得初始页面上的URL列表，在主线程中爬取网页的过程中，不断从当前页面筛选新的URL放入待爬行的URL列表，作为待执行的任务队列，由子线程执行，直到满足系统的停止条件[9]。

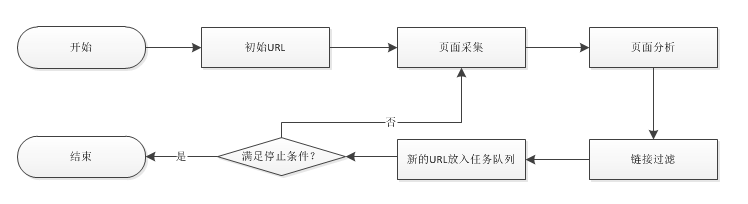


图1 网络爬虫工作流程

其各个部分的主要功能介绍如下：

1. 初始URL：该模块提供初始的URL，以启动爬虫程序。
2. 页面采集模块：该模块是整个网络爬虫系统的入口模块，主要是通过各种web协议（通常为HTTP）来对网页数据进行采集。
3. 页面分析模块：该模块的功能是把页面采集模块采集到的页面数据进行分析，筛选出满足系统需求的URL，加入到任务队列中，如果URL为站点的相对路径，则需要加上网络协议、端口号、域名，使之成为可直接访问的完整的URL。
4. 链接过滤模块：该模块主要是过滤那些满足系统需求的正确格式的链接，并去除已经爬取的重复链接。
   1. 运行环境

运行环境是指代码执行的环境，主要指软件环境，本系统运行于Node.js环境下。

Node.js是一个基于Chrome’s V8 JavaScript engine的JavaScript运行环境（Runtime）。通过这个运行环境，可以让JavaScript像后端语言（Java，Python之类）那样，进行文件、数据库等操作。

Node.js产生的初衷，是Ryan Dahl想设计一个高性能的Web服务器。他发现高性能Web服务器有几个要点：事件驱动，非阻塞I/O。此外，Ryan Dahl设计了Node.js[10]。

Node.js具有以下特点：

Node.js以单进程、单线程的模式运行，它采用事件驱动的的机制实现并行操作。避免了系统分配线程及线程间通信的开销，可以更高效率的利用cpu的性能，降低内存的消耗。

不同于其他编程语言，Node.js采用非阻塞I/O，提高程序性能，更好的提高单个线程的利用率。

Node.js的设计思想是以事件驱动为核心的，Node.js提供的绝大多数API都是基于事件的、异步的风格[11]。事件驱动、异步编程使得程序在执行时无需阻塞等待某个操作（如：HTTP请求、磁盘读写、数据库操作）的完成，使有限的系统资源能得到充分利用。

* 1. 正则表达式

对于网络爬虫系统来说，其实现对网页内容的爬取后，下一步就是利用正则表达式对网页内容进行筛选，提取所需的有效信息。

正则表达式的主要功能是实现字符串的匹配和截取，在爬虫程序中正则表达式对逻辑判断意义重大。本文实现爬虫的过程中，对URL及其他重要字段的解析与判断依靠正则表达式实现。

常见的正则表达式元字符解释如表1：

表1 常用正则表达式元字符

|  |  |
| --- | --- |
| ^ | 匹配一个输入或一行的开头 |
| $ | 匹配一个输入或一行的结尾 |
| \* | 匹配前面元字符0次或多次 |
| + | 匹配前面元字符1次或多次 |
| ? | 匹配前面元字符0次或1次 |
| {n} | 精确匹配n次 |
| {n,} | 匹配n次以上 |
| {n,m} | 匹配n-m次 |
| [xyz] | 匹配这个集合中的任一一个字符(或元字符) |
| [^xyz] | 不匹配这个集合中的任何一个字符 |
| \d | 匹配一个数字字符 |
| \D | 匹配一个非数字字符 |
| \s | 匹配一个空白字符 |
| \S | 匹配一个非空白字符 |

本章对网络引文爬虫相关的概念及技术背景进行了简单的介绍，为后续的系统设计与实现提供了理论基础，具体包括：引文的概念、网络引文的概念、网络爬虫的概念、网络爬虫的工作流程以及网络爬虫用到的基础技术Node.js和正则表达式。

1. 需求分析
   1. 业务需求描述

本系统以万方数据库为例，爬取万方数据库中引用有网络引文的文献，然后将其持久化到数据库中。以《聚焦爬虫系统研究综述》为例，在论文详情页可以看到如下图2内容，点击图中（1）处的参考文献会出现右侧具体的参考文献的列表，如果参考文献较多，则会如（3）处分页显示。在（2）处出现有[EB/OL]标识的即为引用的网络引文，系统首先根据搜索关键字筛选文献列表，再针对每篇文献分页爬取参考文献列表，并解析筛选含有“[EB/OL]”的参考文献，然后将该文献持久化到数据库中。



图2 参考文献网页截图

* 1. 功能性需求分析

1. 网页下载与信息过滤

系统首先对某搜索关键字搜索到的论文页面进行下载，然后根据DOM树分析得到的DOM选择器筛选出需要的论文列表，包括：论文的标题、论文在万方数据库中的访问URL、论文在万方数据库中存储的ID。之后再根据论文的URL爬取论文详情页的参考文献列表部分。

1. 抽取有效数据

根据论文URL爬取论文详情页面的参考文献时，需要分页爬取所有的参考文献页面，并从这些参考文献列表中抽取后缀含有“[EB/OL]”的参考文献，即表示该论文引用了网络引文。最终将这些引用了参考文献的文献持久化到数据库中，以方便后续数据分析。

1. 任务队列

系统需要提供一个或多个拥有一定并发量的任务队列来管理和维护任务，将任务的分发和任务的执行独立开，降低系统耦合度和各模块之间的依赖。

1. 数据的更新与存储

网络爬虫对根据爬取规则下载HTML页面，并根据DOM选择器筛选出所需数据后，需要将获取到的有效数据存储到MySQL数据库中，并根据关键字段进行内容的更新。

本章从系统的业务需求、功能性需求两个方面描述了系统的各方面需求。并具体阐述了网页下载与信息过滤、抽取有效数据、数据的更新与存储这些功能性需求，为下一步网络引文爬虫系统的设计提供了思路。

1. 系统设计方案
   1. 系统结构

根据上一章对系统需求的分析，本章设计网络引文爬虫系统，主要采集引用了网络引文的文献及具体的网络引文标题。该系统应达到以下条件：

1. 易配置性。仅需通过简单的配置即可完成系统参数配置。
2. 爬取速度快。采用任务队列的方式，异步执行任务，充分利用Node.js任务驱动的特性，提高系统资源的利用率和系统吞吐量。
3. 较高的数据准确度。通过正则表达式准备匹配有效信息[12]。

网络引文爬虫系统的系统结构如图3所示：

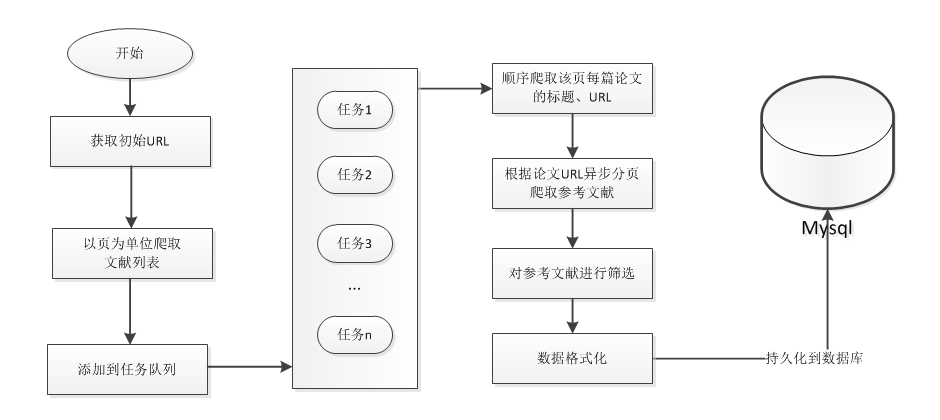


图3 网络爬虫系统结构

* 1. 功能设计

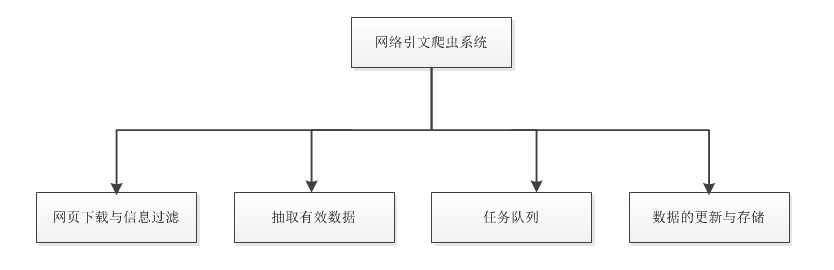
网络引文爬虫系统主要包含了网页下载与信息过滤、抽取有效数据、任务队列、数据的更新与存储四大功能模块，如图4所示：

图4 网络引文爬虫系统功能模块

网页下载与信息过滤模块主要负责发送HTTP请求，下载HTML网页，并直观地过滤掉一些不可能存在有效信息的部分，以避免无意义的操作，浪费系统资源。本系统从初始URL开始下载网页，过滤掉网页中除了文献列表以外的其他网页部分，筛选出文献列表并生成任务添加到任务队列的末尾。

抽取有效信息模块负责对下载的HTML进行有效数据的进一步筛选。系统通过正则表达式调用相应的匹配规则对文献列表中各标签的内容进行匹配，匹配出最大页数以分页爬取文章标题以及判断引文标题是否含有“[EB/OL]”字样等。

任务队列模块负责提供任务执行的环境，该任务队列可根据系统性能合理配置并发量，等待执行的任务进入队列末尾等待执行。这样在提高资源利用率的同时，可方便的监控队列中任务的积压情况。

数据的更新与存储模块负责将爬取到的有效数据存储到MySQL数据库中，MySQL通过唯一索引来达到数据消重的效果。

本章主要从系统总体设计、系统各功能模块设计两个方面详细阐述了系统的结构及主要功能，包括：网页下载与信息过滤、抽取有效数据、任务队列、数据的更新与存储，为下一步系统的具体实现提供了方案。

1. 系统实现
   1. 系统环境

本系统在win7环境下测试运行，具体环境如下表2所示：

表2 系统环境

|  |  |
| --- | --- |
| 硬件环境 | CPU：Intel i7-5500U；内存：8GB DDR3；硬盘：1TB SATA3 |
| 操作系统 | Windows 7企业版 |
| 开发工具 | WebStrom 2017.1  Node.js v4.4.4  MySQL v5.0.95 |

* 1. 网页下载与信息过滤功能

在此模块中，首先通过分析搜索结果页面的网页结构得到初始URL，在图5中（1）处，可以观察到（2）处出现URL，由此可粗略判断该URL为包含文献列表结构的URL页面：

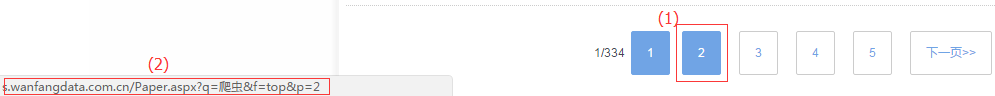


图5 搜索结果页码链接网页截图

在图6中，通过Chrome的开发者工具，我们清楚的看到并验证了图5中（2）处的URL，即：“http://s.wanfangdata.com.cn/Paper.aspx?q=keyword&f=top&p=page”为包含文献列表的初始URL，且从图6中可清晰的看到HTML文档的DOM结构，通过“.left-record .record-title .title”选择器即可对每篇文献的标题部分HTML进行定位及进一步筛选数据。

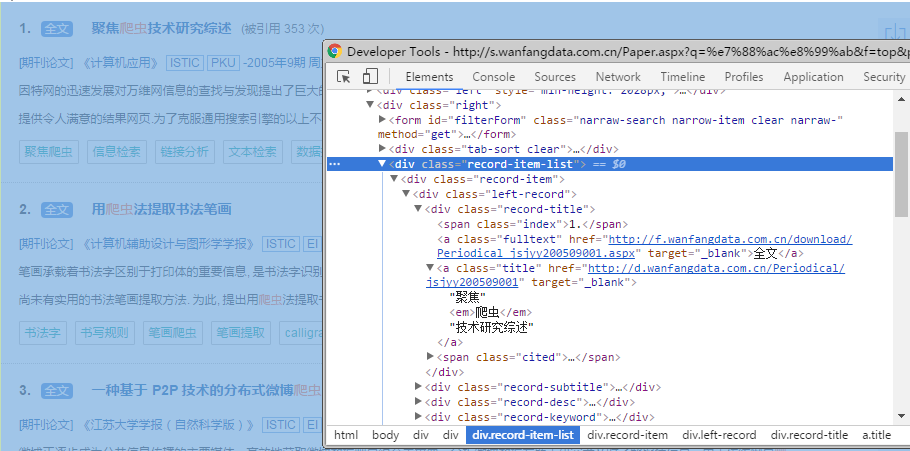


图6 网页DOM结构截图

该模块通过Node.js第三方包axios方便的实现HTTP请求的发送，下载网页内容，接下来通过cheerio对下载的HTML进行过滤，过滤出文献的列表及每篇文献对应的标题、URL、文献ID，提供给数据抽取模块进行下一步的数据抽取。如果在此过程中发生错误或出现非预想的结果，则处理异常并打印异常日志。该部分的两个步骤是爬虫系统的公共步骤，除了具体实现细节不同外，大体逻辑和结构都是一样的。

* 1. 抽取有效数据

在论文的详情页面中，我们需要爬取参考文献列表部分的HTML，但是在详情页面中我们并没有找到对应的HTML。因此，初步判断参考文献列表通过AJAX异步加载。为了验证这一假设，再次打开Chrome工具，保证图7中（2）处Network面板下内容为空，点击图7中（1）处的参考文献，可以看到在图8的Network面板下（1）处出现了一次HTTP请求，在图8中（2）处可以看到该HTTP请求的具体参数，如URL、请求的方法、远程IP等信息，在图8中（3）处出现了参考文献列表，由此我们得到了参考文献的HTTP链接，即：“http://d.wanfangdata.com.cn/CiteRelation/Ref?id=articleId&page=page”。

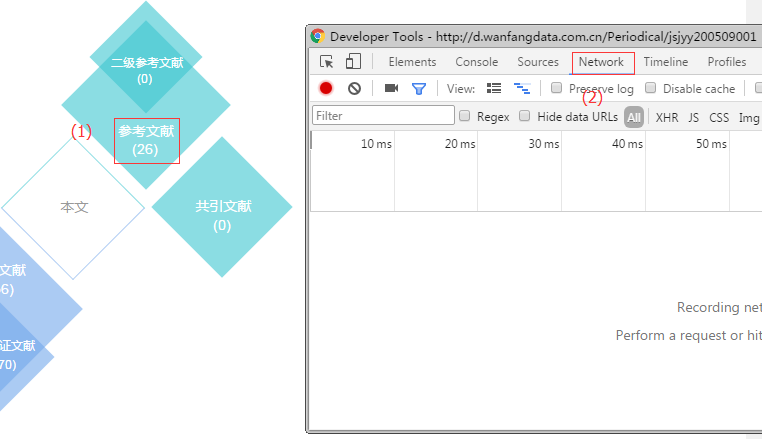


图7 Chrome开发者工具截图1

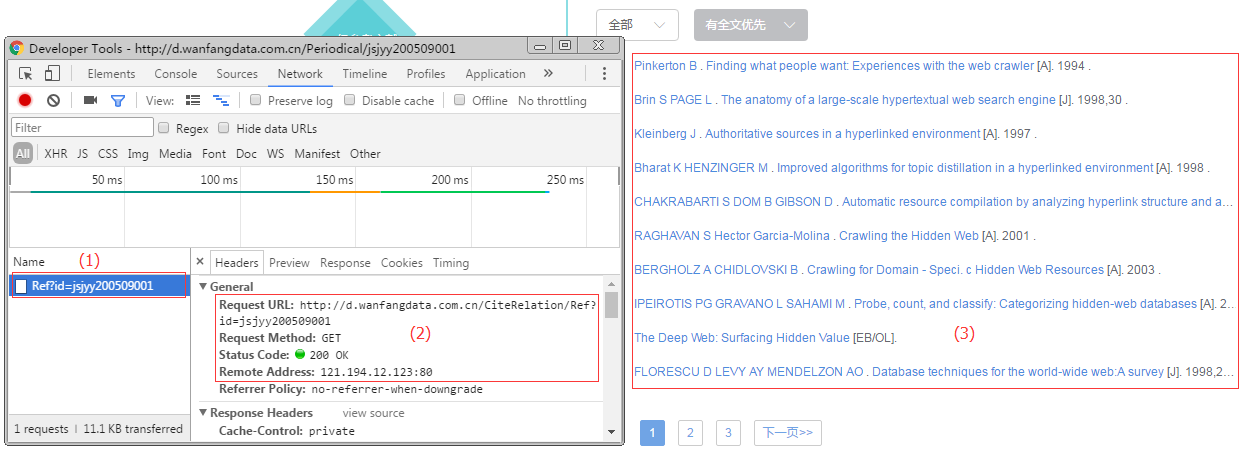


图8 Chrome开发者工具截图2

在该模块实现中，我们依然通过网页下载与信息过滤模块的公共逻辑实现发送HTTP请求，解析HTML的功能。在实现细节上，通过一系列与业务相关的逻辑判断参考文献列表中是否包含“EB/OL”，即网络引文。

* 1. 任务队列

在该模块中，通过async创建了一个可配置并发数量的任务队列，即当并发数为n时，同一时刻最多同时有n个任务在执行，后续任务进来时被加入任务队列末尾和其他待执行任务一起排队等待执行。

* 1. 数据的更新与存储

该模块根据文章标题生成对应的MD5加密字符串，然后通过MySQL数据库连接包mysql实现数据的更新与存储，加密后的md5作为唯一索引用于防重。

1. 系统测试

该部分以“爬虫”关键字搜索到的第一页论文为例，逐篇下载并查看对应的参考文献，通过人工对比的方式验证抓取数据的准确性。

系统配置如下：

|  |
| --- |
| {  // 关键字  keyword: '爬虫',  // 分页并发数  pageConcurrency: 2,  // 单篇论文并发数  itemConcurrency: 5,  // 开始页  startPage: 1,  // 页数  pageCount: 1,  // 响应超时时间(ms)  timeout: 3000  } |

爬取到的数据如下：

|  |
| --- |
| [2017-05-12 09:56:33.138] [INFO] out - { articleTitle: '面向蒙古文主题的网络爬虫采集策略模型',  articleUrl: 'http://d.wanfangdata.com.cn/Periodical/bjgydxxb201507009',  articleId: 'bjgydxxb201507009',  md5: 'b5abfceda1a5902a6e70f4c938a5fbd9',  text: [ '文化中国中国少数民族语言现状', '中国互联网络信息中心第34次中国互联网络发展状况统计报告' ] }  [2017-05-12 09:56:33.325] [INFO] out - { articleTitle: '聚焦爬虫技术研究综述',  articleUrl: 'http://d.wanfangdata.com.cn/Periodical/jsjyy200509001',  articleId: 'jsjyy200509001',  md5: '770ae4654db34c1f9c86f4e6c4f217ca',  text: [ 'The Deep Web: Surfacing Hidden Value' ] }  [2017-05-12 09:56:33.388] [INFO] out - { articleTitle: '基于度排序的P2P IPTV分布式爬虫系统设计与实现',  articleUrl: 'http://d.wanfangdata.com.cn/Periodical/scdxxb-gckx201403016',  articleId: 'scdxxb-gckx201403016',  md5: '06e4dabd1d4339e120822e3a8f1f27df',  text:  [ '上海聚力传媒技术有限公司PPTV homepage',  '上海众源网络有限公司PPStream homepage',  '北京时越网络技术有限公司UUSee homepage' ] }  [2017-05-12 09:56:33.528] [INFO] out - { articleTitle: '聚焦爬虫技术研究综述',  articleUrl: 'http://d.wanfangdata.com.cn/Periodical/jsjyy200509001',  articleId: 'jsjyy200509001',  md5: '770ae4654db34c1f9c86f4e6c4f217ca',  text: [ 'Wu JABERER KUsing SiteRank for Decentralized Computation of Web Document Ranking' ] }  [2017-05-12 09:56:33.637] [INFO] out - { articleTitle: '基于度排序的P2P IPTV分布式爬虫系统设计与实现',  articleUrl: 'http://d.wanfangdata.com.cn/Periodical/scdxxb-gckx201403016',  articleId: 'scdxxb-gckx201403016',  md5: '06e4dabd1d4339e120822e3a8f1f27df',  text: [ 'CNNIC中国互联网络信息中心第32次中国互联网络发展状况调查统计报告' ] } |

在手工下载该页论文后，发现确实存在且仅存在以下网络引文，如下图：

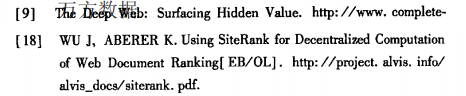


图9 《聚焦爬虫技术研究综述》论文截图

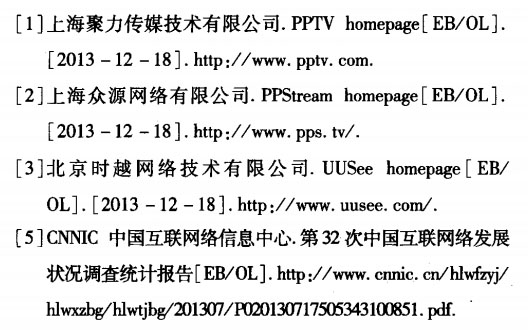


图10 《基于度排序的P2P IPTV分布式爬虫系统设计与实现》论文截图

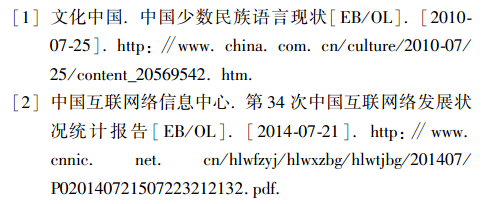


图11 《面向蒙古文主题的网络爬虫采集策略模型》论文截图

* 1. 抽取准确性

爬取数据与手工数据一致，满足数据抽取的需求。测试时，爬取20页论文，并对爬取到的论文随机抽查验证，均通过验证。数据抽取的准确性接近100%。这样的数据准确性为后续的数据分析和数据挖掘提供了高质量、高标准的数据保证，满足了业务需要。

* 1. 抽取用时

网络爬虫程序通常被用来为搜索引擎、数据挖掘提供数据支持。本文通过对万方数据库进行定向爬虫，在爬取速度、数据准确性方面满足了业务需求，测试数据如表3所示（由于受到网络状况及万方服务响应时间的影响，数据仅供参考）：

表3 测试结果

|  |  |
| --- | --- |
| 页数 | 平均用时（S） |
| 1 | 3.395 |
| 5 | 9.068 |
| 10 | 13.056 |

本章主要给出了网络引文爬虫系统的模块关系图，并简要介绍了网络引文爬虫系统各功能模块的具体实现，然后在测试环境下测试数据爬取的速度和准确率，并给出了测试数据。从测试的结果看，系统运行状态良好，能基本满足业务需求。

1. 总结

本文主要针对当前学术环境下关于网络引文的数据分析、数据挖掘不方便的问题，提出了业务需求，并根据需求设计、设计了一套网络引文爬虫系统，该系统涵盖以下功能：

1. 网页下载与信息过滤：系统通过定向数据爬取的方法从万方数据库下载HTML，并初步过滤掉部分网页。
2. 抽取有效数据：系统采用正则表达式对网页中的信息进行精确匹配以抽取有效数据。
3. 数据的更新与存储：系统使用MySQL来存储抽取到的数据。
4. 任务队列：使用任务队列管理爬虫任务，将任务细分，提高系统效率和资源的利用率。

本文给出了该系统的具体实现代码，对系统进行了简单的测试，在最后以表格的形式给出了系统测试的结果（包括：爬取速度、数据准确性），表明了系统运行正常，可以满足业务需求。

由于本人时间和能力有限，系统还有很多问题需要进一步解决，具体问题如下：

1. 系统监控：系统应该有一个监控模块，将系统细分并监控每个细节的性能，如：HTTP请求响应时间、任务队列积压情况等。
2. 日志管理：目前系统日志虽然能满足当前业务需求，但在系统逐渐扩展之后，就需求一个日志模块对系统的日志进行统一管理，如错误日志的处理、失败的HTTP请求的重试策略等。
3. 可视化：目前系统只是在控制台实现操作，没有一个可视化的页面供使用者操作。

致谢

在本次毕业论文即将完成之际，首先我要对我的导师郭强表示感谢，郭老师在论文写作方面给了我很多思路，也给予我了很多帮助。郭老师严谨治学的态度、一丝不苟的工作作风、平易近人的处事风格值得我们每个人学习，谢谢郭老师！

其次，我要感谢在论文写作过程中给予我帮助的各位同学，和你们一起讨论、分析论文是一件快乐的事情，谨在此向你们表示感谢！

还有我的室友、家人一直在我求学期间给予我精神上的支持和物质上的帮助，谢谢你们！

最后要感谢我的母校郑州大学以及信息管理学院对我的培养！

谢谢！

张磊 2017/05/08 于 河南郑州

参考文献

1. 数据挖掘基础知识[EB/OL]. <http://www.xuexila.com/baikezhishi/1244762.html>
2. 李海燕.网络舆情爬虫系统的设计与实现[D].厦门:厦门大学,2014
3. 姜博.基于聚焦爬虫的web信息采集技术研究[D].北京:北京工业大学,2011
4. Boyce D,Kirson A,Schofer J,Design and implementation of ADVANCE[C].IEEE Proceeding of 3rd International Conference on Vehicle Navigation and Information
5. 马晓佳.网络引文分析与传统引文分析的比较[D].南京:南京大学,2011
6. 杨思洛.国外网络引文研究的现状及展望[J].中国图书馆学报,2010,(4)
7. 王建芳.网络引文信息的挖掘及其对数字图书馆资源建设的启示[D].太原:山西大学, 2004
8. 赵茉莉.网络爬虫系统的研究与实现[D].成都:电子科技大学,2013
9. 曾伟辉.支持AJAX的网络爬虫系统设计与实现[D].合肥:中国科学技术大学,2009
10. node.js [EB/OL]. <http://baike.baidu.com/item/node.js?fromtitle=nodejs&fromid=11244313>
11. 王越.基于nodejs的微博系统的设计与实现[D].成都:电子科技大学,2014
12. 陈瑜芳.主题爬虫系统的研究[D].武汉:武汉理工大学,2010