# 基于 HttpClient 与 HTMLParser 的网页正文提取

陈智彬, 崔鸿雁

(北京邮电大学信息与通信工程学院, 北京 100876)

**摘要:** 随着互联网的高速发展,针对互联网的分析处理显得日益重要。本文研究了 HttpClient、HTMLParser等技术,提出并实现了一种基于 HttpClient 与 HTMLParser 的网 页抓取解析方法,该方法能够快速有效对 HTML 页面进行抓取解析,提取出所需的文本内容。

**关键词:** 正文提取; HttpClient; HTMLParser; Hadoop

10 中**图分类号**: TP393.092

5

15

20

# Web Content Extraction based on HttpClient and HTMLParser

Chen Zhibin, Cui Hongyan

(Information and Communication Engineering School, Beijing University of Posts and Telecommunications, Beijing 100876)

**Abstract:** With the rapid development of the Internet, the analysis on the Internet is becoming more and more important. This paper studies the HttpClient, HTMLParser technology, put forward and realizes a web capture and analysis method based on HttpClient and HTMLParser, this method can capture the HTML page and then extract the text content fast and effectively. **Keywords:** text extraction; HttpClient; HTMLParser; Hadoop

#### 0 引言

随着互联网时代信息与数据的快速增长,出于科研、工程和商业等目的对网页数据进行 针对性的提取显得越来越重要。提取网页信息是为后续的进一步处理工作服务的,比如网页 的分类、聚类、关联分析等等。当前,网络所承载的业务类型也越来越多并不断呈现出新的 特点,与此同时,网络上用户行为也呈现出多样化的特点。为了更好的分析用户的网上行为 特点,需要对用户浏览次数频繁的网页进行分析。由于要对海量的网页信息进行处理分析, 不可能采用浏览器人工阅览的方法。因此,通过我们需要编写自己的分析处理程序,提取出 30 有用的信息,为进一步分析做准备。

由于自身固有的缺点,原始 HTML 页面的数据格式并不适合进行处理<sup>[1-2]</sup>。本研究将结合 HttpClient 与 HttpParser 对 HTML 页面进行抓取解析;同时针对大数据量的处理情况,采用基于 Hadoop 的分布式多线程架构。

#### 1 开发工具与环境简介

#### 35 1.1 HttpClient

网页抓取,就是把 URL 地址中指定的网络资源从网络流中读取出来,保存到本地。类似于使用程序模拟 IE 浏览器的功能,把 URL 作为 HTTP 请求的内容发送到服务器端,然后读取服务器端的相应资源<sup>[3]</sup>。虽然 java.net 包中已经提供了访问 HTTP 协议的基本功能,但是对于大部分应用程序来说还不够丰富和灵活。

作者简介:陈智彬,(1987-),男,硕士研究生,主要研究方向:下一代网络、云计算。 通信联系人:崔鸿雁,(1977-),女,副教授,未来网络理论与应用. E-mail: cuihy@bupt.edu.cn

## 中国科技论文在线

40 HttpClient<sup>[4]</sup> 是 Apache Jakarta Common 下的子项目,用来提供高效的、最新的、功能丰富的支持 HTTP 协议的客户端编程工具包,并且它支持 HTTP 协议最新的版本和建议。HttpClient 能够处理 HTTP 连接中的各种问题,使用起来十分方便,只需导入 HttpClient.jar 包,就可以模拟 IE 浏览器来获取网页内容。

#### 1.2 HTMLParser

45

50

55

60

65

HTMLParser<sup>[5]</sup> 是用来解析 HTML 文件的 Java 库,主要用于转换和提取两方面。它能超高速实时解析 HTML 文件,而且不会出错,是目前最好的 HTML 解析和分析的工具。HTMLParser 可以实现文本抽取、链接抽取、资源抽取、链接检查、站点检查、链接重写、网页内容拷贝、内容检验、HTML 信息清洗以及将 HTML 页面转成 XML 格式。

HTMLParser 采用了经典的 Composite 模式,通过 RemarkNode、TextNode、TagNode、AbstractNode 和 Tag 来描述 HTML 页面中的各元素<sup>[3]</sup>。HTMLParser 为我们提供了非常方便的 HTML 解析方式,不同的应用可以采用 Visitor 方式来遍历 HTML 节点提取数据,也可以采用 Filter 方式来过滤节点,提取出我们所关注的节点,再对节点进行处理<sup>[6]</sup>。

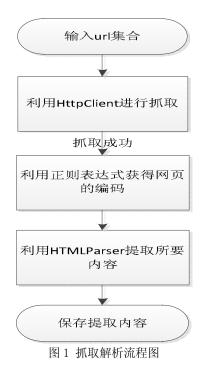
#### 1.3 Hadoop

Hadoop<sup>[7]</sup>是 Apache Lucene 创始人 Doug Cutting 创建的,起源于另外一个开源的网络搜索引擎项目 Apache Nutch,是 MapReduce<sup>[8]</sup> 分布式程序设计模型的开源实现架构,用于在集群上对海量数据进行并行处理。Hadoop 已成为很多大型互联网公司计算架构的核心部分,比如 Google、Yahoo、FaceBook、LinkedIn 以及 Twitter 等。它具有可扩展、经济、可靠、高效等特点,非常适合用于海量数据的分析。

#### 2 系统结构

本系统采用 HttpClient 对输入的 URL 进行抓取,当抓取成功时,则将 HTML 文档保存下来。由于中文网页的编码可能是 UTF-8 或 GB2312 格式的,因此如果没有采用对应的编码格式进行解析的话会产生乱码的问题。据调查 80%以上的网页都没有设置 response Header 的 charset,所以直接通过 HttpClient 获取 charset 是不可靠的;但是有 99%的网页都设置了meta 元素的 charset<sup>[9]</sup>,所以可以先根据一个默认编码(ISO-8859-1)获得 HTML 文档,再根据 HTMLParser 或正则表达式获得该文档的 charset。最后通过使用 HTMLParser 的TitleTag、MetaTag、TextNode 等标签来获得 HTML 文档的文本内容。

整个抓取解析的流程如图 1 所示。为了处理海量的 URL 信息,可以在 Hadoop 平台上 实现此抓取解析应用程序。在分布式多线程的环境下,指向相同主机的 URL 将被分配到同一个抓取队列中。



70

#### 3 系统实现

```
以下代码为实现抓取解析功能的调用方法:
75
          public static ProtocolOutput getHttpOutput(String url) {
               int status = -1;
               HttpClient httpclient = new HttpClient();
               //set http connection timeout
               httpclient.getHttpConnectionManager().getParams().setConnectionTimeout(5000);
80
               GetMethod getMethod = new GetMethod(url);
               //set get request timeout
               getMethod.getParams().setParameter(HttpMethodParams.SO_TIMEOUT, 5000);
               //set retry handler
               getMethod.getParams().setParameter(HttpMethodParams.RETRY_HANDLER,
85
                        new DefaultHttpMethodRetryHandler());
               try {
                    status = httpclient.executeMethod(getMethod);
                    if( status == HttpStatus.SC OK) {
90
                        String html = new String(getMethod.getResponseBody(),"ISO-8859-1");
                        String charSet = getCharSet(html);
                        String input = new String(html.getBytes("ISO-8859-1"), charSet);
                        String content = "";
                        try {
95
                             content = readText(input, charSet);
                         } catch (ParserException e) {
                             e.printStackTrace();
                        URLContent urlContent = new URLContent(content);
```

# 中国科技论文在线

```
100
                       return new ProtocolOutput(urlContent,HttpStatus.SC OK);
                   } else {
                       return new ProtocolOutput(null, status);
               } catch (HttpException e) {
105
                   LOG.error("Please check you provided http address!" + e.getMessage());
               } catch (IOException e) {
                   //network abnormal
                   LOG.error(e.getMessage());
               } finally {
110
                   getMethod.releaseConnection();
               return new ProtocolOutput(null, status);
           }
           其中返回类型 ProtocolOutput 是存储提取内容与响应状态的类。getCharSet 方法采用正
       则表达式获得网页的编码格式,其实现代码为:
115
      public static final String reg= "<meta[^>]*?charset=([a-z|A-Z|0-9]*[\\-]*[0-9]*)[\\s|\\S]*";
      private static String getCharSet(String content) throws PatternSyntaxException {
           Matcher m = pCharSet.matcher(content);
           if(m.find()) {
120
               return m.group(1);
          return "UTF-8";
      }
           网页中的标题、keywords 和 Description 三项对于网页属性的描述具有指导作用,因此
      提取正文也应该将它们考虑在内,readText 实现了对去噪后的网页的正文提取:
125
      private static String readText(String input, String contentEncode) throws ParserException {
               String content = delTag("script",new StringBuffer(input)).toString();
               Parser parser = Parser.createParser(content, contentEncode.trim());
               NodeList nodelist:
130
               NodeFilter textFilter = new NodeClassFilter(TextNode.class);
               NodeFilter titleFilter = new NodeClassFilter(TitleTag.class);
               NodeFilter metaFilter = new NodeClassFilter(MetaTag.class);
               OrFilter lastFilter = new OrFilter();
               lastFilter.setPredicates(new NodeFilter[] { textFilter, titleFilter, metaFilter });
135
               nodelist = parser.parse(lastFilter);
               Node[] nodes = nodelist.toNodeArray();
               .....
           }
           对于不少 HTML 页面, HTML 元素的长度超过了正文的长度, 有时候会混入很多
140
      JavaScript。HtmlParser 经常会将这些元素误认为是正文加以识别,导致很多正文竟然是一段
      JavaScript<sup>[3]</sup> 。delTag 方法的作用就是去除特定标签内容,这里用来去除 JavaScript。
```

通过对百度(UTF-8)、新浪(GB2312)、雅虎(UTF-8)这三家具有不同页面编码的主页进行 抓取解析,发现可以正确得到正文,如图 2 所示:

### 中国科技论文在线

年被灭星围绕的生月太准了30岁以后你有多少身价伸准你未米老公是富蒙0厂分鬼推存近三年你会有什么厄机十二星座12月年底运势抢先看笨蛋测试你敢来吗看看你的占有欲有多强测测你有多好色性成熟度测试女性入口20年后你会胖多少哪种有钱人能带你发财你会嫁给怎样的男人结婚后你会失去什么设为首页关于雅虎帮助中心联系我们诚聘英才商务合作广告服务全部服务警方提示防诈骗远离黄赌毒阿里巴巴集团-阿里巴巴淘宝网一淘网支付宝阿里云淘宝天下天下网商阿里研究中心万网京ICP证000022号增值电信业务经营许可证82-20090145国家药监局京-经营性-2009-0014京网文20110580-195号京公网安备110000000019号广播电视节目制作经营许可证京字第659号Copyright 2011 Yahoocomcn © 版权所有 不得转载使用须知著作权声明不良信息举报中心文明办网文明上网举报如果发现不良信息请您点击此处举报网络110报警服务北京网络行业协会北京互联网违法和不良信息举报热线

http://www.baidu.com 百度一下你就知道百度一下你就知道搜索设置nbspnbsp登录新nbsp闻 网nbsp页 贴nbsp吧 知nbsp道 MP3 图nbsp片 视nbsp频 地nbsp图输入法手写拼音关闭空间 百科 hao123更多gtgt把百度设为主页加入百度推广搜索风云榜关于百度About Baiducopy2011 Baidu使用百度前必读京ICP证030173号

http://www.sina.com.cn 新浪首页新浪首页新浪网为全球用户24小时提供全面及时的中文资讯内容覆盖国内外突发新闻事件体坛赛事娱乐时尚产业资讯实用信息等设有新闻体育娱乐财经科技房产汽车等30多个内容频道同时开设博客视频论坛等自由互动交流空间登录登录名密码选择去向免费邮箱VIP邮箱2008邮箱同名邮箱CN邮箱微博博客视频论坛会员中心您好选择去向免费邮箱微博博客视频论坛会员中心修改密码nbsp安全退出会员中心客服帮助企业邮箱手机新浪网设为主页新浪导航新闻军事社会财经股票基金科技手机数码体育英超NBA娱乐明星音乐汽车报价买车博客微博轻博视频综艺大片房产EJU家居读书教育育儿女性健康中医乐库尚品收藏空间邮箱出国论坛交友SHOW游戏玩玩星座城市广东上海生活旅游电商短信应用手游高尔夫下载导航商城天气爱问彩票公益世园热销创业旅游团购夏日东南亚海岛游推荐广东热门旅游景点导航参加美国之旅活动型投传设中的秦珠八始来情三亚的四天三海里长户山的天池海极日木震后还能去旅游吗美丽险西景区

图 2 正文提取结果

# 4 结论

145

150

本文提出了一种基于 HttpClient 与 HTMLParser 相结合的网页正文抓取解析方法,该方法可以准确快速地提取出用户所需的内容,并且很好地解决了中文乱码的问题。为了应对海量数据的处理情况,在 Hadoop 分布式平台上实现了该方法。

#### [参考文献] (References)

- [1] 洪辉, 刘子敏.智能 Web 信息提取系统的研究和设计[J].微计算机信息,2005,21(11):71-74.
- [2] Valter Crescenzi, Giansalvator Mecca, Paolo Merialdo. An Automatic Data Grabber for Large Web Sites[C]. Proceedings of the 30th VLDB Conference, 2004, 1321-1324.
  - [3] 罗刚, 王振东.自己动手写网络爬虫[M]. 北京: 清华大学出版, 2010.
  - [4] http://hc.apache.org/httpclient-3.x/.
  - $[5]\ http://html parser. source forge.net/.$
  - [6] 牛春山.基于 Lucene 和 HTMLParser 技术的搜索引擎的设计与实现[D].西安:西安电子科技大学,2008.
- 160 [7] http://hadoop.apache.org/.
  - [8] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Communications of the ACM, vol.51, no. 1, pp. 107-113, January 2008.
  - [9] http://hi.baidu.com/sawen21/blog/item/200e9e31985032a25edf0e02.html.