一种批量抽取动态 Web 信息系统①

马 龙[1]，张春涛[1]，杨德仁[2]

[1](宁夏万纬信息技术公司，银川 750000)
[2](宁夏医科大学理学院，银川 750000)

摘 要：针对从 Web 页面获取信息的广泛需求，分析了从中提取信息的关键技术如 URL 地址、HTML 页面和 HtmlParse 解析库；以从 Google Map 中获取企业黄页信息为例，根据从中自动提取数据的技术和步骤，设计和实现了该系统原型，并指出的相关问题及其解决办法。

关键词：Web 页面；HtmlParse；Google 地图；信息抽取；系统

Batch Extraction Information System from Dynamic Web

MA Long[1], ZHANG Chun-Tao[1], YANG De-Ren[2]

[1](Ningxia Wanwei IT Technology Co, Yinchuan 750000, China)

[2](Science College of Ningxia Medical University, Yinchuan 750000, China)

Abstract: In order to respond some extensive requirements for getting information from Web pages, some key techniques such as URL, HTML page and HtmlParse API, were analyzed. Getting yellow page information from Google maps was taken as an example, and according to related techniques and steps of abstracting information from it, the system prototype was designed and implemented. Some related problems were presented, and its corresponding solution were discussed too.

Keywords: Web page; HtmlParse; Google map; information extract; system

1 引言

在互联网时代，人们的工作和生活都与互联网息息相关。但在 Web 上浏览和获取信息时，信息良莠混杂，往往不尽人如意。如何准确提取有价值信息是一件非常棘手问题，相关技术及其应用一直是研究热点。

2 关键技术

在 Web 上发布、检索和提取信息，要涉及到一些相关技术，如 URL 地址编码、HTML 页面和 HtmlParse 解析库等。

2.1 URL 地址与 UTF8 编码

在互联网上，绝大多数信息是通过页面方式发布和展现的。页面文件可以是静态的，也可以是动态的。动态页面的数据来源于数据库。网页文件或以静态文件存在 Web 服务器中，或通过存在于 Web 服务器中的模板与查询数据库的结构数据生成，在浏览器上展示给用户，每个页面都有其唯一地址，即 Web 地址 URL，俗称网址。只要知道网址，就可以浏览、下载和处理页面信息。用户可编写程序下载这类页面并进行相应处理。

静态页面的网址是用户可识别的，编码简单。

动态页面的网址，因待查询字串而异，其数据来源于后台数据库。如在百度搜索“大学”，其地址信息是：

http://www.baidu.com/s?wd=%B4%F3%D1%A7，

其中，%B4%F3%D1%A7 是“大学”的一种编码方式，称为 UTF8 编码。

动态网页的数据规律性很强，可用于批量抽取有用信息。

2.2 HTML 及其标签

大多数网页文件是用超文本标记语言(HTML)语

① 基金项目:宁夏科技攻关计划项目(KGX-01-10-01)
收稿时间:2010-07-16;收到修改稿时间:2010-08-19

言写的。HTML 文本是由 HTML 标签行组成的描述性文本，HTML 标签可以说明文字、图形、动画、声音、表格、链接等。HTML 文件的结构包括头部(Head)、主体(Body)两大部分，其中头部描述浏览器所需的元数据信息，而主体则包含所要说明的具体内容。

2.3 HtmlParse 解析库

HTML 文件是一个半结构化的文档，它以树状结构组织数据，<html>是树根，而有用信息一般含在叶子节点。HTML 网页处理是数据提取的关键环节之一，很容易解析 HTML 代码。

有很多出色工具可用于解析 HTML 代码。其中，HtmlParse 是用纯 java 写的开源 html 解析库，是比较著名并且得到广泛应用的技术之一，可用于改造或提取 HTML 结构或信息。其特点是结构设计巧妙，使用速度快，非常实用，基本满足解析网页的需求。用 HtmlParser 遍历网页内容后，可以得到以树(森林)结构保存的结果。HtmlParser 访问结果内容的方法有两种：使用 Filter 和使用 Visitor。

3 一种自动批量抽取企业黄页信心系统的分析和设计

以从 Google 地图页面中提取企业黄页信息为例，分析手工整理的过程，根据上述核心技术，设计一种从中自动提取黄页信息的系统原型。

3.1 从 Google 地图中提取企业黄页信息过程分析

3.1.1 企业黄页信息

企业黄页信息的基本组成域为：企业名称，行业类别，所在城市，所属城区，街道和电话等。

Google 地图提供了丰富的本地信息服务，可以分类查询本地企业信心。但用手工从 Google 地图页面中提取企业黄页数据时，步骤繁杂、工作量大，如从中提取银川市兴庆区的饭店信息时，其结果多达 20 多页。还要考虑同“饭店”的同义词提取信息，如“饭馆”、“餐厅”、“饭庄”、“食府”等。

3.1.2 手动提取数据过程

手工提取可大致分为：首先，登录 http://ditu.google.cn；其次，输入查询条件，如：银川 兴庆区 医院；则得显示结果的截图如图 1 所示。

图 1　查询医院的信息截图

提取出的这类原始信息，数据量大、耗时，手工处理极其复杂。把这类信息黏贴到文本文件中后，尚待除去链接信息和无用信息。好在动态网页的数据有规律可循，可以编程处理。如何自动提取这种数据，为此提出了一种解决方案及其相应的实现方法。

3.2 系统设计与实现

图 1 中的第 1 条数据，对应 HTML 原文件简化后的代码片段如下：

```
<span id=title class="fn org" dir=ltr>
  银川市金凤区医院
</span>
<span class=adr id=adr dir=ltr>
  宁夏回族自治区银川市庆丰街 152
 </span>
 <span class=tel>
  0951-5036167
</span>
```

HtmlParse 根据 HTML 标签及属性将标签值结构化成一个树状模型，通过遍历这个树就可以获得网页的数据。首先要掌握目标数据属于那个标签，其次，要了解该标签的特殊属性；即要根据标签及其属性就可以过滤数据。采用 HtmlParse 解析以上简化后的代码，得到的结果是：

银川市金凤区医院
宁夏回族自治区银川市庆丰街 152
0951-5036167

该结果数据包含了企业黄页数据的重要内容。

3.2.1 设计思想

系统的基本设计思想是：

首先，根据查询字串，写出 URL;

其次，根据 URL 获得承载数据的页面；

再者，利用用 HtmlParse 解析页面文件，并保存结果。

3.2.2 Htmlparser 解析过程分析

利用 Htmlparser 对 html 源文件进行解析，Htmlparser 把 html 解析成具有树状结构的数据模型，并对整个树做遍历，再通过特定的过滤器将不必要的数据过滤掉。

在上面例子的处理中，第一个条件是，Htmlparser 通过条件过滤掉一部分信息；其次是以的 class 属性为判断条件；过滤后得到真正需要的数据。

图 2 是 Htmlparser 的解析过程：

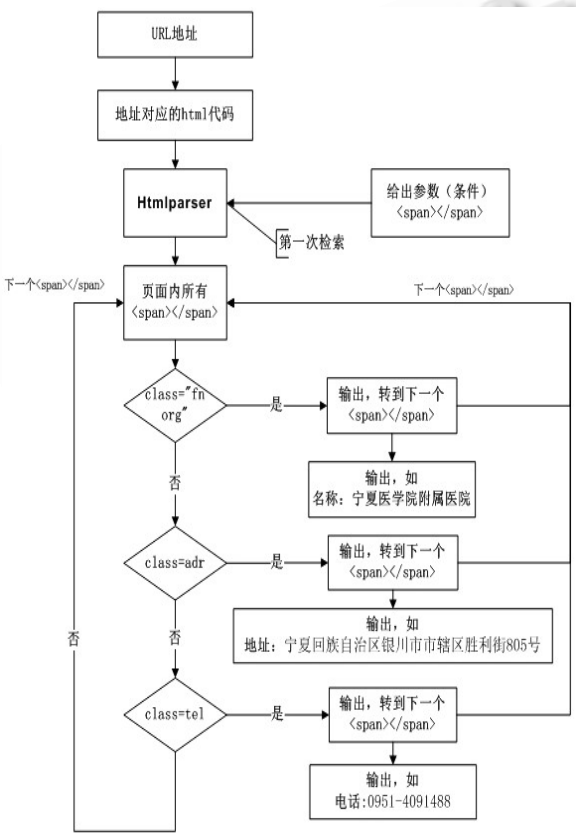

图 2 Htmlparser 解析过程

3.2.3 一种自动提取企业黄页信息系统原型设计与实施

① 参数设计

运行前，要准备“行政区划分文档”，作为参数，如：北京市、上海市等；准备“行业分类文档”，要提取那些行业的数据，如：教育、餐饮、医院等，行业文档还有一个作用，解决了同义词问题，将所有可能的同义词都编入这个文档，程序运行时就能读到这些数据。

为以上两个参数文档做接口，以便采集人员修改。系统对用户提供两种接口，第一，参数设定，参数包括区域信息(如：西夏区、金凤区)、行业信息(医疗、教育)，程序执行之前先把这些将要用到的参数准备就绪。

② 程序处理流程

程序运行过程如下：

a) 读取区域文件

b) 读行业分类数据

c) 用以上得到的两个参数生成 URL 地址，用该地址获取网页源文件

d) 用 HtmlParse 解析网页源代码，将得到的数据保存到数据库

e) 判断行业数据是否为最后一个

如果是，执行 f)

如果不是，就去读下一个行业数据，执行 b)

f) 判断是否为最后一个区域

如果是就退出程序，至此程序运行完毕。

如果不是,去读下一个区域，执行 a)

图 3 是程序流程图：

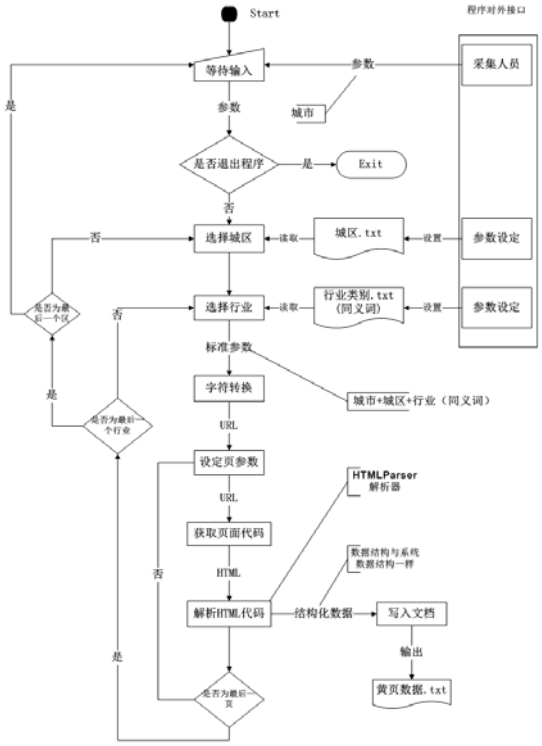

图 3 系统流程图

③ 系统主界面

系统由 Java 语言开发，为了能更好的分类，数据的提取以城区为单位，系统运行结果将保存在数据库。系统主界面如图 5 所示。

图 4　系统运行界面

④ 待完善之处

对 Google Maps 深入分析发现，通过程序查询到的数据是以当地政府为中心的一个有限区域的企业数据，这个区域是以当地政为中心的周边地区，用一个距离参数来表示，如图 1 所示。

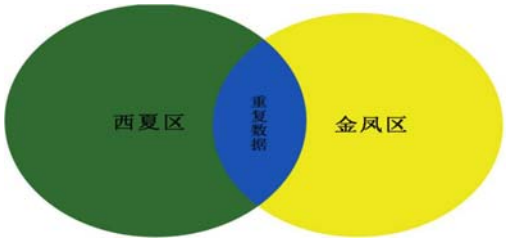

图 5　重复数据示意图

本系统虽然能比较完整的提取所需的数据，但依然存在一些待完善的地方，如结果中有大量重复数据存在。

图 5 形象地表示银川的两个城区的重复数据问题，这个问题是可以解决的，或在写入数据库时先判断当前数据是否存在，如果存在就舍弃本次插入操作；或者在提取的数据中本身就有一个表示距离的参数，通过限定距离来提取数据也能避免重复数据的存在，不过这样做可能会遗漏部分数据。

4　结语

在信息化时代，必然离不开互联网这个共享网络，如何合理、高效地利用这些共享资源是值得研究的课题。本文以从 Google 地图中提取企业黄页信息为例，简述了提取动态页面信息的过程，设计和实现了相应的系统原型，并对系统中待完善之处提出了解决方案。

参考文献

1 http://htmlparser.sourceforge.net/
2 http://ditu.google.cn/maps?hl=zh-CN&tab=wl
3 http://allenj2ee.javaeye.com/blog/222454
4https://www.ibm.com/developerworks/cn/opensource/os-cn-crawler/

(上接第 35 页)

当系统架构复杂，涉及的特征和影响因素繁杂时需要变更划分方法中的基本单元划分和相关度分析影响因素。

参考文献

1 杨孟辉,廖建新,吴乃星.下一代网络核心业务平台的可靠性分析.通信学报,2006,(4):60－66.
2 祁卓娅,王建正,韩新民.模块柔性划分方法.机械工程学报,2007,43(1):87－94.
3 史俊友,陶庆斌,翟红岩.基于遗传算法的产品族模块划分.青岛科技大学学报,2010,(4):188－193.
4 王日君,张进生等.基于公理设计与模糊树图的集成式模块划分方法.农业机械学报,2009,(4):179－183.
5 廖安舟,王纯.移动广告系统的研究与设计.计算机系统应用,2009,18(8):15－18.
6 OSGi Alliance, OSGi Service Platform Core Specification Release 4, April 2007. http://www.osgi.org