基于事件驱动的车型参数主题爬虫①

张会福，周亚平

(湖南科技大学 计算机科学与工程学院，湘潭 410128)

摘　要：网络上的大量数据都隐藏在深层网络中，普通的主题爬虫只能抓取表层数据，而对深层数据的抓取则力不从心。通过建立模拟事件驱动模型，采用基于链接的 BM 改进过滤算法和基于向量空间模型的内容过滤算法，结合汽车车型参数主题，利用 Htmlparser 解析动态生成的页面，对其进行结构化处理后将数据存储于索引数据库中，由此实现车型信息的自动抽取和解析。实验结果表明，该系统模型针对同领域数据具有良好的事件触发适应性和高过滤准确率。

关键词：事件驱动；主题爬虫；网络爬虫；车型参数；Ajax

Topic Crawler of Car Parameters Based on Event Driven

ZHANG Hui-Fu, ZHOU Ya-Ping

(Department of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 410128, China)

Abstract: Large amounts of data on the Internet were hidden in the deep Web. Traditional topic crawler could only get surface data, and crawl the deep data was powerless. This paper established simulation model of Event-Driven, Used improved filtering algorithm based on BM by links, and content filtering algorithm based on VSM(vector space model), with model parameters of car, used Htmlparser to analysis dynamically generated pages, processed their structures to store in database, thus achieve extraction and parsing of car's information. The results of experimental show that the model for the same field datas has a good event trigger adaptability and high filtering accuracy.

Key words: event-driven; topic crawler; crawler; car parameters; Ajax

网络爬虫是一个自动从网上抓取网页的程序。主题爬虫[1]是在网络爬虫的基础上加入爬行条件，针对某一个特定的主题，再通过过滤算法、分类算法和计算与主题相似度[2]来过滤与主题不相关的链接和内容，再建立索引数据库，方便后面的检索工作。

用通用的网络爬虫与深层网络主题爬虫进行比较，可将网络上的数据分为两类：表层数据和深层数据[3]。表层数据就是传统爬虫能够抓取到的数据，主要是指通过某个网页源文件上的静态链接能够导向的网页数据；深层数据是指不能通网页上的链接可访问到的页面，主要有网页上要通过触发表单事件后动态生成链接所链接的网页和动态生成的网页数据。

动态数据的产生主要是因为 JavaScript 技术和 Ajax 技术在网页中的使用。由于 Ajax 技术的盛行，大部分数据呈非结构化或半结构化形式，传统的网络爬虫不会解析 JavaScript 已不适应当前发展的需要，于是诞生了支持事件驱动[4]的爬虫，基于事件驱动能解决异步调用并解析异步回调逻辑和内容，获得动态 DOM 语义结构产生的内容。而智能的主题爬虫又能合理利用网络带宽和硬件资源，大大减少了对不相关主题网页的抓取。然而在汽车行业，尤其是大型门户网站，编辑对汽车车型的参数设置之多而感到束手无策，如果人工输入车型参数再提交将浪费大量的时间。本文提出的支持事件驱动的车型参数主题爬虫能够智能地解决此问题。

① 基金项目:国家自然科学基金(50775070);湖南省科技厅项目(2009FJ4055);湖南省教育教育厅项目(10K023)
收稿时间:2011-03-06;收到修改稿时间:2011-03-30

1 事件驱动及Ajax技术

Ajax 是一种创建交互式应用的网页开发技术，它使用 JavaScript 操作 DOM 进行动态显示及交互；并使用 XMLHttpRequest 对象与 Web 服务器进行异步数据交换。它与传统的协议驱动不同的是，传统的 web 是通过在表单输入数据提交之后，向服务器发出请求，服务器端接收到请求和数据之后，再导向另外一个网页。它们对 JavaScript 是缺乏语义理解的，不能再去模拟触发事件的异步调用，而且它们默认每个页面的 DOM 结构是静态的。而基于 Ajax 的事件驱动机制是使用一些基于 XML 的 Web service 接口，并在客户端采用 JavaScript 处理来自服务器的响应，从而许多工作在客户端完成，减少了等待服务器响应的时间，速度更快，节约网络带宽。 目前伴随着以 Ajax 驱动为代表的 Web 应用的大规模流行，一系列巨大的新挑战也开始加速降临到目前网络爬虫上，新的爬虫机制的建立已迫在眉睫。Ajax 应用颠覆了以往基于纯 HTTP 协议驱动的爬虫机制，即默认所有的页面资源都是直接由超链接所指向的，现有的爬虫只需模拟用户的超链接请求并解析对应的响应页面，然后再分析页面内容、语义以及衍生的超链接，以此进行爬网。

2 基于事件驱动的车型主题爬虫模型

基于事件驱动的车型主题爬虫(Event-Driven of Car Topic Crawler,EDCTC)总体分成五部分：事件驱动模拟模块、网页获取模块、JavaScript 解析模块、过滤模块、页面解析模块。系统总体框图如图 1 所示：

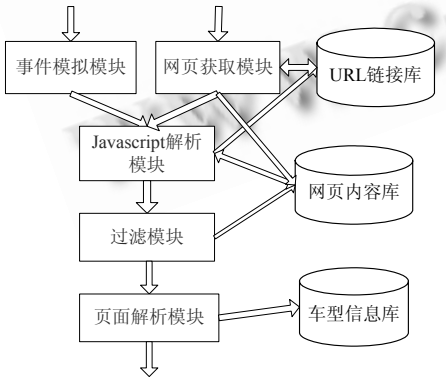

图 1 EDCTC 框图

事件驱动模拟模块：以车型信息查询表单的共同特征为基础，模拟车型的选择后提交。

网页获取模块：用来抽取从某汽车网开始抓取网页，获得 URL 和 JavaScript 脚本。

JavaScript 解析模块：以从事件触发所响应的 JavaScript 脚本和 Html 页面解析出的 JavaScript 文件作为被解析的对象，调用了 JS 作为脚本引擎来解析，同时结合 spidermonkey 解析器[5]来辅助解析，根据解析出的数据映射到对应的标签，形成一个新的动态网页。

过滤模块：运用改进的基于 BM 过滤算法[6]去过滤不相关的链接和非车型信息页面的链接。

页面解析模块：首先对 Html 页面进行规范处理，再采用 Htmlparser 解析器[7]结合正则表达式解析出车型参数配置信息，将数据存入车型信息库。

3 EDCTC的设计

3.1 事件驱动模拟模块

此模块主要通过模拟表单提交并触发 JavaScript 事件，获得新产生的动态链接。综合所有汽车网站上的车型查询表单的共同特征，具体如何模拟事件驱动以 www.dgecar.com 首页的车型查询表单为例：首先通过抓取此页面的源文件后，分析含有<form>、<select>、<option>标签，获得这些标签中的内容存入数据表中，并使内容与内容形成对应关系，如一个品牌对应多个车型，一个车型又对应多个型号。为了使此模块不产生错误的对应关系（如奥迪，华晨宝马新 3 系，宝马 320i 豪华型），此模块在获取车型时会判断车型属于哪个品牌，型号又对应哪个车型。然后根据每一条对应记录分别触发一个事件，从而生成动态页面新 URL。

3.2 网页获取模块

从给定的汽车主题相关的 URL 开始抽取，通过 HTTP 协议自动获取 Web 页面和 JavaScript 脚本文件。此模块采用了宽度优先策略[8]，抓取到每个网页上面的所有链接和 JS 文件的相对路径，存入队列中，使用先进先出[9]方式抓取队列中 Url 对应的网页源文件存放到网页内容库中。另外一部分网页是来自于 JavaScript 解析后得到的动态页面。以循环从 Url 队列中获得网址去抓取网页内容，实现全网的页面抓取。

3.3 JavaScript 解析模块

用 spidermonkey 解析器分析 JS 脚本，寻找异步请求函数 XMLHttpRequest，如有此函数，继续执行异步调用函数，获得返回的数据，刷新 DOM 树。另外要解析出 JavaScript 事件产生的新链接。例如根据如下代码片段

从 JavaScript 中解析出品牌，车系，车型的对应关系：

```
var b307='$众泰汽车,s3554|众泰梦迪博朗,s1630|众泰 2008,s3009|新众泰 2008,s3434|众泰 5008';
function GetCarXiLie(ob,obj)
{
var pinpaiId=ob.value;
switch(pinpaiId)
{
var menu = b60.split('$');
for(var i=1;i<menu.length;i++)
{
var ary1 = manu[i].toString().split(',');
for(var j=1;j<ary1.length;j++)
{
var ary2=ary1[j].toString().split('|');
obj.options.add(newOption(ary2[1].toString(),ary2[0].toString()));
}
……
}
```

同时解析出提交后的新链接。

3.4 过滤模块

3.4.1 构建汽车车型参数配置的向量空间模型

根据某汽车网所有的车型信息页面的共同特征，建立如表 1 所示的向量空间模型，在系统中将根据该向量空间模型来作为解析 Html 页面和过滤非车型页面的条件。

表 1　车型参数的向量空间模型表

车型	车身重量	轴距	轮距	全车长度	…	标准引擎
奥迪 A1	1	1	1	1	1	1
比亚迪 F0	0	1	0	0	1	0
大众 Polo	1	1	1	1	1	1
…	…	…	…	…	…	…

由以上向量空间模型可知，奥迪 A1 后面对应的参数都为 1 可知此页面正是车型参数配置页面，这些参数特征词的相似度[9]要大于程序给定的相似 0.9，并将它的参数信息存入数据库的 Car 表中。其中从比亚迪 F0 对应的参数配置的相似度可得知此页面不是车型详细参数页面。

3.4.2 过滤算法

过滤算法是对 web 信息进行聚类和分类[10]，以获得用户感兴趣的信息，而减少对无关主题信息的抓取，节约硬件资源，提高爬虫的准确率。本文用到的有对链接 BM 过滤和对内容过滤两个算法。

① 基于链接的 BM 过滤算法

BM(Boyer-Moore)算法的思想是：字符串的匹配顺序是从右到左，匹配最右边的字符成功，则循环不断向左匹配；采用字符串滑移启发式策略。

此算法根据大型汽车网的车型链接的共同特征，与列出的所有汽车车型在一个页面上，然后通过每款车型而链向其参数配置。在以下四个方面进行了过滤：

(a) 过滤掉网站域名来参与匹配；

(b) 过滤掉链接后面紧接时间的链接，这些一般为汽车新闻；

(c) 过滤掉链接中字符串过长的链接；

(d) 过滤掉链接中字符个数小于 30 的链接。

② 基于向量空间模型的内容过滤算法

根据汽车车型参数配置网页页面的特征[10]，它不同于新闻页面，新闻页面文字比较多，而参数配置页面一般表格比较多，数字比较多，但是时间类型的数据比较少，一个<td></td>中字符串的长度比较短，meta 中的 keywords 中一般含有“参数”、“配置”、“汽车”、“车型”等字符串。另外再通过向量空间模型中给定的特征词，计算相似度，如果相似度超过 0.9，可知该网页是某一款车型的参数配置页面，否则将网页过滤。

3.5 页面解析模块

首先对 Html 页面进行规范处理，由于所有的车型页面的源文件代码都有一个共同特征：就是用表格标签来将各个参数信息列出来，只需将源文件映射成 DOM 树结构，再采用 Htmlparser 解析器结合正则表达式解析出车型参数配置信息，将数据存入车型信息库。

4　测试结果与分析

表 2　用事件驱动技术获得的动态网页数据表

网站	动态网址数	动态页面下载数	丢失
dgecar.com	4404	4404	0
pcauto.com.cn	7204	7204	0

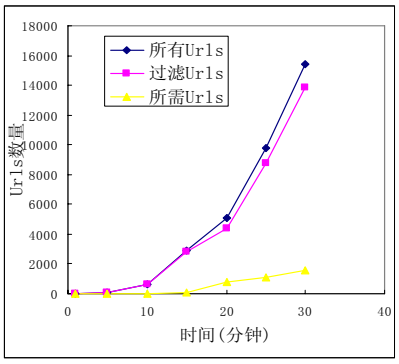

图 4 系统爬取到某汽车网的 URLs 截图

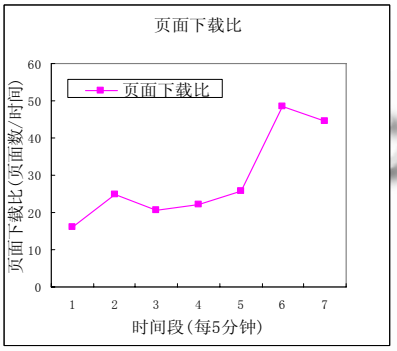

图 5 EDCTC 模型下载页面的响应时间比例曲线

图 4 显示了通过给定的某大型汽车网的 URL 所抓取到的总链接数、过滤掉的链接数、留下的链接数。从图中可以观察出该爬虫在刚开始的十多分钟爬到的网页基本上都不是车型信息页面，因此这些页面虽然都与汽车相关，但不是需要的页面。大概在二十分钟后，分析出是车型的页面逐渐增多.另外可以观察到数据过滤的准确率高。图 5 显示的是每隔五分钟下载的页面数比例，其中页面下载比由以下公式计算：

页面下载比 = 下载页面总数/当前消耗时间(min)

从图 5 可知，在不同的时间段下载的页面数是不定的，它由多方面因素来决定的，有页面数据的大小，网速，网络带宽，以及动态页面的生成与解析时间等等。表 2 中的数据显示了不同站点利用事件驱动技术对车型查询表单的抽取所获得的动态车型网址以及下载的动态页面数，丢失未被下载的动态车型页面为 0。

由此可知 EDCTC 系统的事件触发模型的适用性良好，且适合各种不同领域的同一类型特征网页的抓取和解析。

5 结论

本文研究了支持事件驱动的车型参数主题爬虫的设计与实现，陈述了它的整个框架的设计，在支持事件驱动方面，建立了模拟事件触发模型，采用了 JavaScript 解析器查找 XMLHttpRequest 异步请求返回的动态数据。用车型参数配置的向量空间模型，来指定特征词即车型配置参数，用其去匹配网页内容和过滤网页。同时还结合各大型门户汽车网车型参数页面特征，对爬取到的链接和内容进行过滤，从而既减少了爬虫的工作量，又提高了信息过滤的准确率。

参考文献

1 周立柱,林玲.聚焦爬虫技术研究综述.计算机应用,2005,25(9):1965–1969.

2 刘金红,陆余良.主题网络爬虫研究综述.计算机应用研究,2007,24(10):26–29.

3 曾伟辉,李淼.深层网络爬虫研究综述.计算机系统应用,2008,17(5):122–125.

4 Alvarez M, Raposo J, Pan A, Cacheda F, Bellas F, Carneiro V. Deep Bot: A focused Crawler for Accessing Hidden Web Content. ACM, 2007. 18–25.

5 EI-Desoudy AI, Ali HA, EI-Ghamrawy SM. An Automatic Label Extraction Technique for Domain: Specific Hidden Web Crawling, IEEE on Computer Engineering and System, 2006,26(5):454–459.

6 袁小节.基于协议驱动与事件驱动的综合聚焦爬虫研究与实现[硕士毕业论文].长沙:国防科技大学,2009.

7 Kumar M, Vig R. Design of Core: context ontology rule enhanced focused web crawler.International Conference on Advances in Computing, Communication and Control Proc., New York, NY: ACM, 2009: 494–497.

8 王辉,刘艳威,左万利.使用分类器自动发现特定领域的深度网入口.软件学报,2008,19(2):246–256.

9 Somboonviwat K. Simulation Study of Language Specific Web Crawling, Institute of Industrial Science, University of Tokyo. 2005(5):57–62.

10 Pant G, Srinivasan P. Link Contexts in Classifier-Guided Topical Crawlers. IEEE Trans. on Knowledge and data Engineering, 2006,18(1):107–122.