基于高级搜索页面的动态表单搜索①

李海滨，许南山

(北京化工大学 信息科学与技术学院，北京 100029)

摘　要：根据表单项前的文字信息反映表单项输入信息的特点，提出通过解析表单项动态填充表单的方法，解决了应用网站自身高级搜索页面对同一类型的多个网站进行搜索的问题。针对图书类的网站进行研究，利用动态解析表单获得结果页面，对其进行解析并加权排序，最后按照统一的显示格式展现。根据实验结果验证了算法设计的正确性，可利用本算法对多个同类型的网站借助其自身搜索进行搜索查询。

关键词：表单解析；动态填充；结果页面解析；结果项排序

Advanced Search Page-Based Dynamic Form Search

LI Hai-Bin, XU Nan-Shan

(College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract: According to the characteristics that the input information can be reflected by the textual content ahead of input item, this paper has proposed a method to dynamically fill the form items by parsing them, and solved the problem that how to search multiple web sites with the same type using their own advanced search page. In this paper, book shopping website was chosen as research object. Using the method of dynamically form, some resulting pages can be generated, then it will parse and sort these pages by their weights, and at last display them in the same format. At last, a satisfying result has obtained in the experiment and confirmed the algorithm accuracy, thus it can be used to search multiple websites with the same type, using their own search engine.

Key words: form parser; fill dynamically form; parse result page; result item sort

随着互联网的飞速发展，网上信息量大量增加，多数网站为了满足人们的搜索信息需求，推出符合网站业务模式含有多个搜索项的高级搜索页面[1]。主流的传统搜索网站对于单一查询项进行分词，拆分成某一个或多个关键字，通过索引库查询到网络爬虫抓取相关的网页。这种方式忽略了网站自身的高级搜索页面的作用，获取的信息具有大量冗余，而且缺乏针对性。因此利用网站自身的高级搜索功能，对高级搜索页面的表单进行动态填充，对结果关联度进行排序而获得的信息往往比主流传统搜索网站更具有针对性[2-3]。本文针对图书销售网站进行研究，提出一种对多数图书销售网站都适用的动态搜索的方式。

1　动态搜索的流程

如动态搜索流程图所示(见图 1)，在动态搜索流程中，主要流程有由表单动态填充、结果页面解析、结果项排序三部分组成。表单动态填充的主要工作是通过对候选图书网站的图书搜索页面进行 Form 表单解析，从中选择最合适的 Form 表单进行填充，提交表单。结果页面解析将返回的结果页面按照特定的格式进行解析。结果项排序将多个网站的解析结果按照相关度进行排序，将排序后的结果按照统一的格式进行展示。

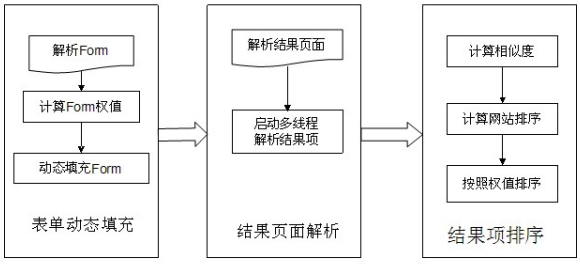

图 1　系统总体框图

① 收稿时间:2011-03-02;收到修改稿时间:2011-04-08

2 动态搜索的算法设计

2.1 表单动态填充

表单动态填充是通过对候选图书网站的图书搜索页面进行 Form 表单解析，从一个或多个 Form 选择最合适的 Form 表单进行填充。对于多个 Form 表单的情况，系统会对每个 Form 表单建立一个线程来解析表单，根据字典计算权值，比较 Form 表单的权值，选出权值最大的 Form 表单进行填充，对于命中的字典项动态修改其命中次数，利于字典中命中率较高的关键字能够迅速进行匹配。字典的预定义是通过对多个图书网站采取人工的方式进行关键字的收集。如果选出的权值最大的 Form 表单的权值仍小于 Form 表单正确解析的阀值，则抛出异常，以邮件的方式通知管理员。管理员通过查看邮件中的异常信息判断并修改字典中的关键字[4]。

文本输入框是 Form 表单中最常用的标签之一，也是 Form 解析的重点。对于文本框的解析采用前向解析的方式[5]，将解析出的关键词与 Form 标签项词典的单词项进行匹配，若匹配成功，将此 Form 的权值加 1，提取出文本输入框的 name 属性，以备进行表单填充。

前向匹配是对一个 HTML 标签之前的文本标签或 Label 标签进行解析，这种文本标签节点或者 Label 标签节点可能位于兄长节点（如图 2 所示），也可能位于文本标签节点或者 Label 标签节点的祖先节点的兄长节点的子孙节点（如图 3 所示）。因此对于这两种不同位置都要进行解析，将两种解析出来的非空结果与 Form 标签项词典的单词项进行匹配取得权值，取权值最高的关键字作为该 HTML 标签的关键字。

```
<div>
    <span>
        <label id="bookNameLbl" for="bookName">Book Name :</lable>
    </span>
    <input id="bookName" name="bookName" type="text" size="20"/>
</dvi>
```

图 2 Label 标签是输入标签的兄长节点示意图

```
<div>
    <label id="bookNameLbl" for="bookName">Book Name :</lable>
</div>
<div>
    <span>
        <input id="bookName" name="bookName" type="text" size="20"/>
    </span>
<div>
```

图 3 Label 标签是输入标签的祖先的兄长的子节点示意图

单选框和多选框需要额外的描述信息来确定此项的表意信息，用户根据单选框和多选框后续的额外描述信息进行选择。下拉框标签中的选项的子节点文本信息具有一定的表意性，但有时需要后向的辅助信息一起进行表意。如图 4 所示，用户根据第一个下拉框后续显示的文字信息“年”，明确第一个下拉框应该填写年份信息。因此用户对于单选框或者多选框的操作通过其后向节点的文本信息进行。因此只对下拉框、单选框和多选框等标签进行前向解析是远远不够的，还要对其进行后向解析。

图 4 后向解析举例示意图

后向匹配是对下拉框、单选框和多选框等标签进行关键词匹配，对标签匹配的方向是该标签的后向，即对一个 HTML 标签之后的文本标签或 Label 标签进行解析，这种文本标签节点或者 Label 标签节点可能位于弟弟节点。

当一个网站所有处理 Form 的多线程全部解析完毕时，比较每个 Form 权值，取权值最大的 Form 进行表单填充。利用之前解析 Form 表单输入项的 name 属性，与 Form 表单的 action 属性联合生成 Form 表单的原生 URL，将用户提交的查询参数进行编码，替代原生 URL 中的输入参数生成实际提交的查询 URL。

2.2 结果页面解析

对于查询结果的展示，各个图书网站采用各自的展现风格进行展示。因此需要对图书网站的展示页面进行整合，提取出公共数据项，统一查询结果的显示格式。统一的结果显示信息，即一个图书信息对象，包含图书的标题、出版社、出版时间、编著者、定价、网站价格、封面图片、描述信息等信息，其中显示标题的位置存在超链接。在显示模块中，点击标题就可以弹出显示该书籍信息的网站网页。

对于结果的解析需要人工的干预，需要预先了解并熟悉图书网站的搜索结果的展示页面的 HTML 标签结构，将这种标签结构进行抽象提取，利用抽象提取出来的标签进行解析获得图书信息对象的链表，完成结果解析的前半部分的操作。

人工预定义的方式是采用 XML 文件预定各图书网站搜索解析的格式，XML 的定义格式如图 5 所示。

website 标签定义图书网站，当结果解析程序解析到 website 标签时会建立一个多线程进行该图书网站的解析子程序。url 标签记录高级搜索页面的 url，通过其与搜索页面产生的结果建立一对一的映射关系。tagParser 标签的子孙节点记录需要解析的节点的信息，根据其子孙节点携带的信息便可以解析出搜索结果。childTag 标签表示该标签的子孙节点。tag 标签记录需要解析的节点的详尽信息，tag 的 name 属性记录该标签的名称，子标签 uniqueAttribute 记录一个或多个该标签的唯一属性，通过这些唯一属性在其父标签的范围内可以唯一确定该标签。子标签 readAttribute 记录一个或多个该标签需要输出的信息，主要应用于提取超链接的 href 属性以定位图书详细信息显示页面和图书封面图片的链接地址。tag 标签的 relatedInfo 属性记录 tag 映射 HTML 标签的文本信息对应哪种图书显示信息。tag 标签的 isLoop 属性是建立多线程解析结果的标志。当程序检测到 tag 标签的 isLoop 属性为 true，程序对于该 tag 的子节点采用多线程的方式进行解析以提高解析效率。

```
<resultParser>
  <website>
    <url>url</url>
    <tagParser>
      <childTag>
        <tag name="BODY">
          <childTag>
            <tag name="DIV">
              <childTag>
                <tag name="LI">
                  <childTag>
                    <tag name="A" relatedInfo="Title">
                      <uniqueAttribute>
                        <attribute name="name" value="p_name"/>
                      </uniqueAttribute>
                      <readAttribute>
                        <attribute isLink="true" name="name" value="href"/>
                      </readAttribute>
                    </tag>
                    <tag name="A" relatedInfo="Author">
                    <tag name="A" relatedInfo="PublishCompany">
                  </childTag>
                  <uniqueAttribute>
                    <attribute name="class" value="maintitle"/>
                  </uniqueAttribute>
                  <readAttribute>
                  </readAttribute>
                </tag>
              </childTag>
              <uniqueAttribute>
                <readAttribute>
                </readAttribute>
              </tag>
            </childTag>
            <uniqueAttribute>
              <readAttribute>
              </readAttribute>
            </tag>
          </childTag>
        </tagParser>
      </website>
</resultParser>
```

图 5　结果页面解析预定义示意图

每个 tag 子线程都会返回一个结果集对象，因此每个 Website 线程可以获得一组结果集对象。如果获得结果集对象组为空，将会抛出异常，主程序捕获异常，将该异常信息以邮件的方式放送给管理员。异常产生的主要原因是图书网站搜索结果页面的 HTML 布局发生变化导致解析程序无法正常解析。管理员接收到异常邮件，及时调整结果解析的 XML 定义文件即可正常解析。

当一个网站所有处理 tag 子线程全部结束时，每个 tag 子线程会返回一个图书信息对象，可以获得一个图书信息对象的链表，作为结果排序项的输入。

2.3 结果项排序

对于结果页面解析出的结果项进行排序，主要考虑的因素是该类似图书在不同网站的出现频数和在各个网站的排序顺序。类似图书的计算方式将一个网站的图书与其他的网站的图书以图书标题的方式计算相似度，计算相似度的依据是图书的图书名，计算公式如式(1)所示（tL 为标题匹配字符长度，lL 为该网站图书标题长度，oL 为其他网站的图书标题长度，m 为其他网站图书总数）。

$$\text{Similarity}=\frac{\sum_{i=1}^{m}\frac{\text{tL}_i}{\sqrt{\text{lL}_i\times oL_i}}}{m} \tag{1}$$

图书在每个网站的排序顺序直观地反映该图书的对查询项的匹配程度和受欢迎程度，因此对排序结果项有至关重要的作用。由于图书相似度与图书在网站排序同等重要，由式(1)可知，图书相似度小于等于 1，因此图书在网站排序的权值应小于等于 1。网站排序的权值的计算采用排序加权法[6-7]，计算公式如式(2)所示（n 为在结果页面的排序数，m 为结果项）。

$$Sortvalue=\frac{m-n+1}{m} \tag{2}$$

图书排序的权值为图书相似度权值和页面排序权值之和，将结果项按照图书排序的权值进行逆序排列，按照统一格式展现给用户。

3　实验结果分析

最后在实验室对算法进行了实验分析，实验硬件环境是 dell 台式机一台，奔腾 4 处理器，512M 内存，XP 系统，网络带宽 100.0Mbps。开发语言是 java，开发环境是 eclipse。作为实验数据源，可以成功解析的图书网站如表 1 所示。

表 1 作为实验数据的图书网站表

89	高级搜索页面 URL
当当网	http://search.dangdang.com/AdvanceSearch/AdvanceSearch.aspx?c=0
China-Pub	http://www.china-pub.com/search/search_index.asp
卓越网	http://www.amazon.cn/mn/article?pageName=book&pageIdent=adv_sr_book
北发图书网	http://book.beifabook.com/Customer/Searchbooks.aspx
华人图书网	http://www.huarenbook.com/search.asp

运行程序进入查询界面，输入相应查询项进行搜索查询，最后返回排序后的统一显示界面。表 2 列出以上五个网站正确解析的结果数。

表 2 正确解析结果数

网站名称	正确解析结果数目
当当网	25
China-Pub	20
卓越网	12
北发图书网	25
华人图书网	12

现有的动态表单搜索技术主要针对输入项的 id 或者 name 属性与字典中的关键词进行匹配，对匹配成功的输入项进行表单提交。这些技术相对于本文中的程序解析速度更快，但一些网站的 HTML 编写具有很大的随意性，输入项的 id 或者 name 属性未必可以与字典中的关键字匹配成功，因此本文中的程序比同类算法搜索查询更精确、冗余率更小。

在实验过程中发现，对于通过 JavaScript 的方式使用隐藏域进行表单提交的网站无法进行表单自动填充。本文不对 JavaScript 解析进行讨论，原因有三点：

(1) JavaScript 是动态脚本语言，可以未对变量进行定义即可使用，因此对于变量的检测有一定困难；

(2) 现今 JavaScript 构架使用广泛，流行的 JavaScript 构架数量较大，而且构架之间的差异性极大，需要对众多框架应用策略模式进行针对性开发，开发工作量较大[8]；

(3) 当今的门户网站（图书网站也在其中）对于常用页面都采用页面静态化的方式提高并发效率[9]，同时遵照一般 Form 表单设计方式进行实现，便于网络爬虫的抓取。

应用 JavaScript 的方式使用隐藏域进行表单提交的图书网站只有一个，因此不对此个体网站进行研究。

对于图书所连接的图片需要代理方式访问的图书封面图片无法展示，其解决方法是利用该网站的 URL 地址作为代理地址以获取封面图片。由于 Socket 属于占用系统资源较多的对象，因此将每个网站建立一个 Socket 对象以通过代理的方式获取封面图片。

4 结语

本文的创新点是：实现网站的搜索页面的表单动态填充，基于表单项前的文字信息反映表单项的所要输入项的特点，提出一种利用网站已有的搜索页面，解析该页面的表单计算表单项权值的策略，实现表单的动态填充，同时解析结果对其进行加权排序。

本文提出一种新的思路，对于同一类型的多个网站，充分利用网站自身搜索引擎进行针对该类型网站的搜索，相对于传统搜索网站，搜索到的信息更具体，更有针对性。根据其他不同类型的网站建立相应的字典，便可以对其类型的网站进行搜索。

参考文献

1 黄晓冬.Invisible Web 研究综述.情报科学,2004,22(9):1144–1148.

2 刘伟,孟小峰,孟卫一.Deep Web 数据集成研究综述.计算机学报,2007,30(9):1475–1489.

3 Chang K, He B, Li C, Patel M, Zhang Z. Structured database on the Web: Observations and implications. SIGMOD Record, 2004,33(3):61–70.

4 马军,宋玲,韩晓晖,闫泼.基于网页上下文的 Deep web 数据库分类.软件学报,2008,19(2):267–274.

5 寇月,申德荣,李冬,聂铁铮.一种基于语义及统计分析的 Deep Web 实体识别机制.软件学报,2008,19(2):194–208.

6 吕蓬,史丽超.层次分析法中排序权值计算的目标规划模型.科技信息,2007,(22):84–85.

7 金菊良,魏一鸣,付强,丁晶.计算层次分析法中排序权值的加速遗传算法.系统工程理论与实践,2002,(11):39–43.

8 金晓鸥.基于 Rhino 的 JavaScript 动态页面解析研究与实现.计算机技术与发展,2008,(2):1–4,50.

9 贺理,吴健,贾彦民.基于 JavaScript 的浏览器端调用 Web 服务研究与实现.中国科学院研究生院学报,2007,(6):801–804.