# 一种高效Web数据抽取包装器的设计与实现

李宏伟,史培中,张素智
(郑州轻工业学院 计算机与通信工程学院,河南 郑州 450002)

**摘 要**:Web包装器是根据特定的抽取规则从特定的Web数据源执行数据抽取程序,设计Web包装器是Web信息抽取和集成的关键技术。详细阐述了一种基于预定义模式的Web包装器的设计与实现过程,并选取了几个出版社的新书发布Web页面进行了数据抽取验证和抽取结果分析,取得了较好的效果。充分体现了此Web包装器的可行性、高效性及可维护性,能够应用在基于Wrapper/Mediator方法的Web数据集成。

**关键词**:包装器;抽取规则;信息抽取;Web数据集成
**中图分类号**:TP311 **文献标识码**:A **文章编号**:1673-629X(2009)02-0123-04

# Design and Implementation of an Efficient Wrapper for Web Data Extraction

LI Hong-wei, SHI Pei-zhong, ZHANG Su-zhi
(College of Computer and Communication Engineering,
Zhengzhou University of Light Industry, Zhengzhou 450002, China)

**Abstract**: Web Wrapper extracts the data from the given Web sources according to the corresponding extraction rules of them. design is a key technology for Web information extraction and integration. Describes the design and implementation of a kind of the Web Wrapper which based on pre-defined schema. Then validates the data extraction from the new books information Web pages of some publishing companies and analyses the extraction results with this kind of Web Wrapper. Find it can accurately extract the data from the Web source. So can conclude that this kind of Web Wrapper which proposed in this paper is feasible, efficient and maintainable. It will be applied for Web data integration based on Wrapper/Mediator.

**Key words**: wrapper; extraction rule; information extraction; Web data integration

## 0 引 言

随着Internet和Web技术的迅速发展,Web页面成为信息发布的主要载体。基于HTML的网页着重描述信息的表现形式,便于用户通过浏览器访问,但它的信息组织方式不便于程序进行加工处理。

Web数据特点[1]:

(1)Web数据不是由任何一个部门或组织所控制的,它来源于各种组织或个人,因而是没有固定的数据模型,即使表示同一语义所使用的数据类型也各异;

(2)Web数据的组织也是任意的,只要能够在Web上展现,满足用户需求即可;

(3)Web数据的内容和表现方式又是动态变化的。

因此Web数据源的数据集成性非常差,给Web应用的建立造成了极大的困难。Web包装器的任务就是负责抽取HTML格式的数据并转化为结构化的数据。基于Web包装器的应用程序能以访问数据库中信息的方式来访问Web数据,所以Web包装器是Web数据集成体系结构中的关键部分,是基于中介模式的Wrapper/Mediator方法的Web数据集成的基础。在研究包装器概念的基础上,采用基于预定义模式的Web包装器的设计思想,综合了算法设计和过程设计,并通过新书发布页面信息抽取实验,分析了算法和系统的性能。

## 1 包装器的概念设计

定义 Wrapper。

给定一个包含一系列Web页面 $P$(其中,$P=\{p_1,p_2,\cdots,p_n\}$)的Web数据源 $S$,找到一个映射关系

**收稿日期**:2008-05-20
**基金项目**:河南省自然科学基金资助项目(0411010500);校博士基金项目(2004-010)
**作者简介**:李宏伟(1966-),女,河南孟州人,助理研究员,研究方向为计算机应用技术、教学方法研究;张素智,副教授,博士,研究方向为Web数据库、分布式计算和异构系统集成。

$W$,它可以将 $S$ 中的 $P$ 映射到一个数据集 $R$,并且当 $p_i(i \in \{1,\cdots,n\})$ 结构变化不大的情况下能正确抽取数据。映射 $W$ 就是通常所说的 Web 包装器(Wrapper)。

从功能上来说,Wrapper 就是根据特定的抽取规则从特定的半结构化 Web 数据源执行数据抽取的程序[2]。Wrapper 的核心是抽取规则。抽取规则是用于从每个 HTML 文档中抽取相关信息。

从生成来说,由于各个数据源的 Wrapper 是要分别建立的,Wrapper 又与 HTML 文档的格式密切相关。如果页面的设计者更改了原有文档的格式,那么 Wrapper 就可能因为无法抽取数据而失效。为了保证连续而正确的 Web 数据抽取,这就需要对 Wrapper 进行维护。根据 Knoblock 提出的 Wrapper 生命周期[3](如图 1 所示),Wrapper 要经过生成、运行、失效后维护、再生成、运行这样一个循环往复的过程。

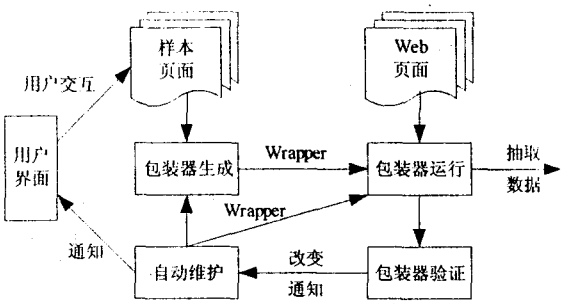


图 1 Wrapper 生命周期

维护的步骤首先要进行 Wrapper 的验证,然后,进入维护过程。当页面发生变化时,Wrapper 所抽取的数据就可能不正确或抽取不到数据,这样就触发了维护例程。维护在本质上是在新页面中重新建立抽取规则,从而完成 Wrapper 的自动修复过程。

人们研究快速有效地自动生成 Wrapper 的目的之一也是减少维护的代价。在这种情况下,一旦 Wrapper 失效,维护工作就变成重新生成 Wrapper。因此,需要探索新的方法来解决 Wrapper 的维护问题。为了简化包装器生成过程,提高包装器对动态页面变化的自适应能力,详细阐述了一种基于预定义模式的 Web 包装器的设计与实现过程。

## 2 Web 包装器的设计

### 2.1 算法设计

Web 信息抽取就是从 Web 页面所包含的无结构或半结构的信息中识别用户感兴趣的数据,并将其转化为结构和语义更为清晰的格式(XML、关系数据、面向对象的数据等)。信息抽取可以理解为一个从待处理文本中抽取信息,形成结构化的数据并存入一个数据库,供用户查询和使用的过程[4]。

为了完成信息的抽取和转化,Web 包装器需要具有四个方面的能力:

(1)信息定位:确定所需要的信息在文档中的位置;

(2)数据抽取:将文本内容分字段抽取出数据;

(3)数据组织:将抽取的数据按照正确的结构和语义组织起来;

(4)可维护性:当 Web 页面发生变化时,Web 包装器仍能正确抽取数据。

因此,设计了一种高效的 Web 包装器算法如下:

**输入:**

- Config. xml 配置文件:Web 数据源 S 抽取规则定义;
- S:Web 数据源;
- P:Web 数据源 S 的 Web 页面,其中 $P = \{p_1, p_2, \cdots, p_n\}$;
- T:HTML 解析后生成的 DOM 树,其中 $T = \{t_1, \cdots, t_n\}$;
- B:待抽取信息块,其中 $B = \{b_1, \cdots, b_m\}$;
- Express:表达式;

**输出:**

- R:抽取数据结果集 $R = \{R_1 \cup R_2 \cdots \cup R_n\}$

① 利用 JDOM 解析 Config. xml 配置;

② $R = \emptyset$(空数据集);

③ for(int i = 1; i <= n; i ++)

{

解析 S 中的 $p_i$ 得到 $t_i$,即:$p_i \rightarrow t_i$

从 $t_i$ 定位信息抽取块 $b_j$,即:$t_i \rightarrow b_j$,其中 $j \in \{1,\cdots,m\}$

//对于 $p_i$ 中得到的 $b_j$ 进行如下操作;

④ for(int j = 1; j <= m; j ++)

{

用表达式 Express 从 $b_j$ 中析取数据,记作 $R_{ij} = \{r_{j1}, \cdots, r_{jk}\}$

$k$ 表示从 $S$ 中抽取数据生成 $k$ 个字段的数据模型

}

⑤ Return $R_i = R_{i1} \cup R_{i2} \cdots \cup R_{im}$,其中 $i \in \{1,\cdots,n\}$

}

⑥ Return $R = R_1 \cup R_2 \cdots \cup R_n$

### 2.2 Web 包装器的设计

根据以上算法,Web 包装器的结构如图 2 所示,主要由 3 个模块组成:预定义模块、数据抽取模块和数据组织模块。其中预定义模块、数据抽取模块是 Web 包

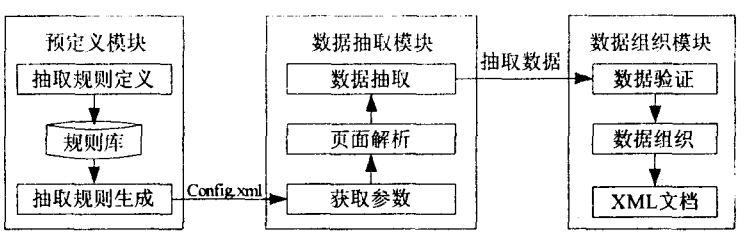


图 2 Web 包装器结构

装器的核心部分。

(1)预定义模块。

预定义模块主要完成了抽取规则定义。文中设计的 Web 包装器是基于规则的抽取模型[5],考虑到这种包装器的可维护性和重用性,采用了通过解析 XML 配置文件(Config. xml)来完成信息定位和信息抽取。对于 Web 数据源页面发生了变动,则 Web 包装器的维护只需要更改针对此 Web 数据源的配置文件(Config. xml)即可。在网页组织形式变动不大的情况下,可以方便、快速地解决了 Web 包装器的维护问题。预定义抽取规则 Config. xml 配置文件模板如下:

```
<? xml version="1.0" encoding="gb2312"? >
<config>
  <url>Web 源网页地址</url>
  <beginPage>起始页</beginPage>
  <endPage>结束页</endPage>
  <tag>标签</tag>
  <index>索引号</index>
  <regex>正则表达式</regex>
</config>
```

(2)数据抽取模块。

数据抽取模块作为 Web 包装器的核心部分,完成了信息定位和信息抽取功能。页面解析主要是解析 HTML 文档格式的文件,可以利用 HTMLParser 解析器。HTMLParser 是一个纯的 Java 写的 Html 解析的库,不依赖于其它的 Java 库,是目前很好的 HTML 解析和分析的工具,能根据需要抓取网页数据或改造 HTML 的内容。此模块主要完成抽取信息的定位,即确定所需抽取的信息块在文档中的位置。

在完成信息定位后,根据定义的抽取规则中的正则表达式来按字段析取出所需要的数据。正则表达式(Regular expression)是一种可以用于模式匹配和替换的强有力的工具[6]。一个正则表达式就是由普通的字符以及特殊字符(称为元字符)组成的文字模式,它描述在查找文字主体时待匹配的一个或多个字符串。正则表达式作为一个模板,将某个字符模式与所搜索的字符串进行匹配,从而可以按字段析取出所需要的数据。

(3)数据组织模块。

Web 包装器完成的功能就是从半结构化的信息中抽取出结构化的数据并保存。因此,如何将抽取的数据以结构化的形式保存起来也是 Web 包装器的一个关键部分。而数据组织模块正是完成了对抽取结果的处理过程。使用 XML 格式来组织抽取结果,XML 语言具有良好的数据存储格式、可扩展性、高度结构化以及便于与数据库交互[7],可以方便以后对抽取信息的进一步处理,如检索和分类等。

## 3　实验和结果分析

为了验证文中提出的 Web 包装器设计与实现方法的可行性、高效性及在 Web 信息集成中的应用,现针对清华大学、冶金工业、北京大学等出版社的新书推荐和发布网站的页面运用以上设计思想定义其抽取规则及 Web 包装器,进行图书信息抽取测试。用 ce, te 和 fe 来分别表示:已抽取出的正确信息个数、还没抽取出的正确信息个数和抽取出的错误信息个数;用 $R$ 表示召回率也称查全率;$P$ 表示精度也称查准率,根据 $R$、$P$ 计算公式[8] 得实验结果如表 1 所示。

表 1　图书信息抽取实验结果

| 出版社 | 网页地址(起始页 - 结束页) | ce | te | fe | $R$(%) | $P$(%) |
|---|---|---|---|---|---|---|
| 清华大学 | http://www.tup.com.cn/book/menu-jp.asp(1~60) | 1175 | 21 | 0 | 98.2 | 100 |
| 北京大学 | http://cbs.pku.edu.cn/scrp/bookcustomore.cfm?iSno=(1~3) | 13 | 0 | 0 | 100 | 100 |
| 冶金工业 | http://www.cnmip.com.cn/xskd.asp(1~136) | 2160 | 1 | 0 | 99.9 | 100 |
| 人民邮电 | http://www.ptpress.com.cn/pptphnews/monthboard.asp | 133 | 0 | 0 | 100 | 100 |
| 电子工业 | http://www.phei.com.cn/(1~513) | 4103 | 0 | 0 | 100 | 100 |
| 机械工业 | http://www.cmpbook.com/stackroom.php(新书上架 1~16) | 306 | 0 | 0 | 100 | 100 |

从以上表中实验结果看出,此包装器的召回率(查

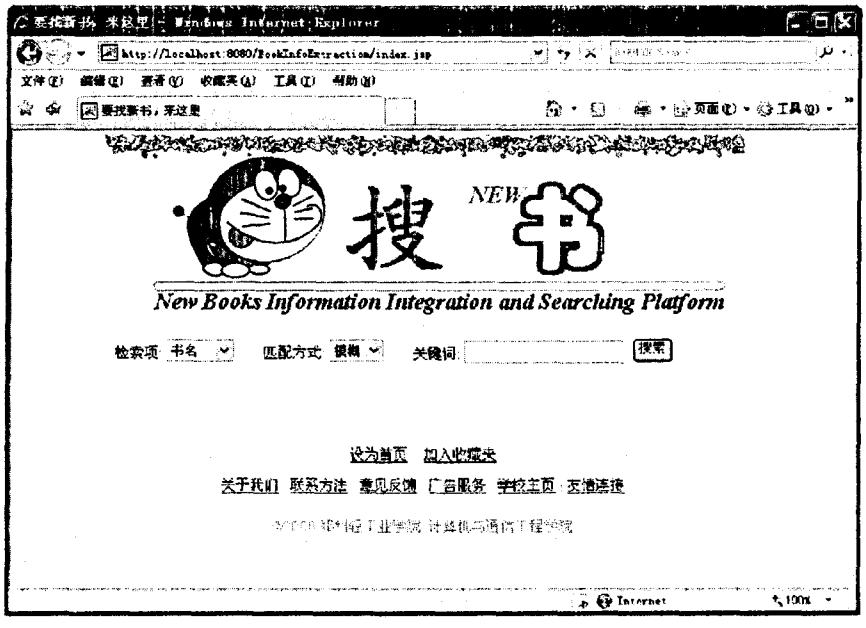


图 3　新书发布信息集成查询系统(部分截图)

全率)和精度(查准率)均可达到近 100%。经过分析,清华大学出版社由于其图书发布页面以列表形式显示。

为了抽取其详细数据,利用二次抽取完成数据抽取,在抽取过程中一些图书信息定位不同造成一些图书信息未能抽取出来。但从整体实验结果来看,此 Web 包装器的设计的可行性,体现了其高效性,可以应用于 Wrapper/Mediator 方法的 Web 数据集成[9]。

## 4 结束语

设计的基于预定义模式的 Web 包装器考虑到了 Wrapper 的维护问题,当 Web 页面发生变动时,只需修复抽取规则(Config.xml)即可,而不用改动、重新编写程序,便于 Wrapper 的维护和源程序的重用。基于此 Web 包装器的设计与实现,已经实现了 Wrapper/Mediator 方法的新书信息集成查询系统(如图 3 所示),通过统一的查询接口,提供新书查询检索服务。

下一步的研究工作是利用此 Web 包装器的设计思想,在不损失其召回率和精度的情况下,设计并实现针对某个领域的通用 Web 包装器,利用 Wrapper/Mediator 方法开发基于 Web Service 的 Web 数据集成系统[10]。


参考文献:

[1] 孟小峰.Web 信息集成技术研究[J].计算机应用与软件,2003,20(11):32-36.

[2] Raposo J,Pan A,Alvarez M,et al.Automatically maintaining wrappers for semi-structured web sources[J].Data & Knowledge Engineering,2007,61:331-358.

[3] Knoblock C A,Lerman K,Minton S,et al.Accurately and Reliably Extracting Data from the Web: A Machine Learning Approach[J].Bulletin of the IEEE Computer Society Technical Committee on Data Engineering,2000,23(4):33-41.

[4] 刘 迁,焦 慧,贾惠波.信息抽取技术的发展现状及构建方法的研究[J].计算机应用研究,2007,24(7):6-9.

[5] 王敬普,林亚平,周顺先,等.基于包装器模型的文本信息抽取[J].计算机应用,2007,27(3):655-658.

[6] 杜冬梅,许彩欣,苏 健.浅谈正则表达式在 web 系统中的应用[J].计算机系统应用,2007(8):87-90.

[7] Holzner S.XML 完全探索[M].北京:中国青年出版社,2001:7-10.

[8] 贺令亚,柳佳刚.基于 Web 的包装器技术的现状与发展[J].电脑开发与应用,2007,20(6):27-29.

[9] 师雪霖,牛振东,宋瀚涛.基于中介器/包装器的联合数字图书馆集成信息检索机制[J].计算机应用,2005,25(3):703-705.

[10] 张素智,李宏伟,李树凯.基于 Web 服务的数据集成[J].郑州轻工业学院学报,2005,20(4):34-37.


(上接第 122 页)

表 1 文中算法与文献[4]的算法结果对比

| 长度 | 文中算法 | 文献[4]算法 |
|---|---|---|
| 3 | 1→8→11→16 | 1→8→11→16 |
| 4 | 1→8→11→14→16 | 1→8→11→14→16 |
| 5 | 1→3→6→10→11→16 | 1→3→6→13→14→16 |
| 5 | 1→3→6→13→14→16 | 1→3→6→10→11→16 |
| 6 | 1→3→6→10→11→14→16 | 1→2→4→9→12→14→16 |
| 6 | 1→3→6→13→14→11→16 | 1→2→3→6→10→11→16 |
| 6 | 1→2→3→6→10→11→16 | 1→3→6→13→14→11→16 |
| 6 | 1→8→11→15→12→14→16 | 1→2→3→6→13→14→16 |
| 6 | 1→2→3→6→13→14→16 | 1→3→6→10→11→14→16 |
| 6 | 1→2→4→9→12→14→16 | 1→8→11→15→12→14→16 |

表 1 列出了两者求解节点 1 到节点 16 的前 10 条最短路径的结果。两者结果相同,不同的只是求得相同长度的路径的顺序不同。

文中算法的优势在于 G1 中路径的唯一性,以至于 G1 中的最短路径一次派生背离路径,就会得到下一条最短路径。而 G1 中的最短路径派生背离路径时,如果其路径前缀很长,那么派生的次数明显减少,时间复杂度将远低于 $O(e \times n^2)$,从而加速了搜索速度。文中算法已经成功应用到中兴通讯基金资助项目"传输网络规划系统 NetNumen™TOP"中,取得了很好的效果。

## 5 结束语

基于背离路径设计的算法,有效地解决了前 Kth 最短路径的搜索问题,其时间复杂度为 $O(e \times n^2)$。在实时应用中,文中的算法有很好的应用前景。


参考文献:

[1] 李乐民.WDM 光传送网的选路和波长分配算法[J].中兴通讯技术,2001,6:4-7.

[2] Mokhtar A, Azizogou M. Adaptive wavelength routing in all-optical networks[J]. IEEE/ACM Trans. on Networking, 1998,6(2):197-206.

[3] Alanyali M, Ayanoglu E. Provisioning algorithms for WDM optical networks[J]. IEEE/ACM Trans. on Networking, 1999,7(5):767-778.

[4] 柴登峰,张登荣.前 N 条最短路径问题的算法及应用[J].浙江大学学报,2002,36(5):531-534.

[5] 王明中,谢剑英,陈应麟.一种新的 Kth 最短路径搜索算法[J].计算机工程与应用,2004(30):49-50.