

复杂 Web 页的 Wrapper 自动化生成技术研究

陈洪平, 方 巍, 李 林, 崔志明

(苏州大学 智能信息处理及应用研究所, 江苏 苏州 215006)

摘 要: 针对基于模板生成 Web 页的基本特点, 结合 Ontology 知识, 探索 Deep Web 垂直搜索中, 复杂 Web 页面的 Wrapper 自动化生成的解决方案. 对实际复杂 Web 页面抽取的实验结果表明, 该方法达到了较高的抽取准确率.

关键词: 本体; 数据抽取; 包装器

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2010)04-0062-04

The Study of Automatic Generation of Wrapper for Data Extraction from Complex Web

CHEN Hong-ping, FANG Wei, LI Lin, CUI Zhi-ming

(Institute of Intelligent Information Processing and Application, Soochow University, Suzhou 215006, China)

Abstract: According to characteristic of the template-based Web-pages, combining the ontology, this paper proposes a method to solve the problem of automatic generation of wrapper for complex Web pages in the deep Web vertical-search condition. Experimental evaluation on a number of real complex Web page collections indicates that our method can gain a high extraction precision.

Key words: ontology; data extraction; wrapper

1 引言

随着万维网的迅速发展, Web 成为巨大的信息仓库. 对 Web 数据进行大规模的抽取、管理及集成, 成为 Web 数据挖掘的重要课题. 由于 Web 数据源的异构性以及弱结构性的特点, 使得对 Web 数据进行大规模信息集成, 成为一个难点. 从互连网中进行大规模、自动化地抽取数据是 Web 大规模数据集成的必要步骤.

近年来, 基于数据导向 (data-intensive) 类型站点的特点, 提出了 Wrapper 的方法^[1-2], 寻求解决此类网站的数据抽取问题. Wrapper 就是被赋予信息抽取任务的计算机程序, 是信息集成系统中的一个组件. Wrapper 的任务就是采用一系列规则, 负责将用户所关心的信息, 从 Web 数据中抽取出来. 由于

Html 文档格式表现各异, 不同的 Html 文档需要不同的抽取规则. 因此, Wrapper 跟特定数据源的网页格式密切相关. 目前使用的 Wrapper 主要存在一些不足. 文中 针对性提出 Wrapper 自动生成的方法.

2 复杂页面的 Wrapper 自动化生成技术

2.1 问题定义

文献[2]中对于面向网页 Wrapper 的研究仍然局限在简单结果页面的抽取上, 如图 1 中的 A_1 所示; 而对类似于图 1 中的 B_1 、 B_2 的复杂页面, 则涉足很少. 通过观察, 我们可以发现复杂页面主要有以下一些特征^[3-4]:

(1) 从产生方式上看, 复杂页面是由网页模板所生成.

(2) 从内容上看, 复杂页面中的数据记录不仅包

收稿日期: 2009-02-14; 修回日期: 2009-04-13

基金项目: 国家自然科学基金项目(60673092); 江苏省重大科技支撑与自主创新项目(BE2008044); 江苏省“六大人才高峰”项目(06-E-037)

括图像、也包括普通文本和锚文本。

(3)从网页布局结构上看,复杂页面数据记录的内容可以被组织成多列或多个区域,或者同一模板

下的数据记录布局可以根据条件而不同.如图2中 B_1 、 B_2 所示。

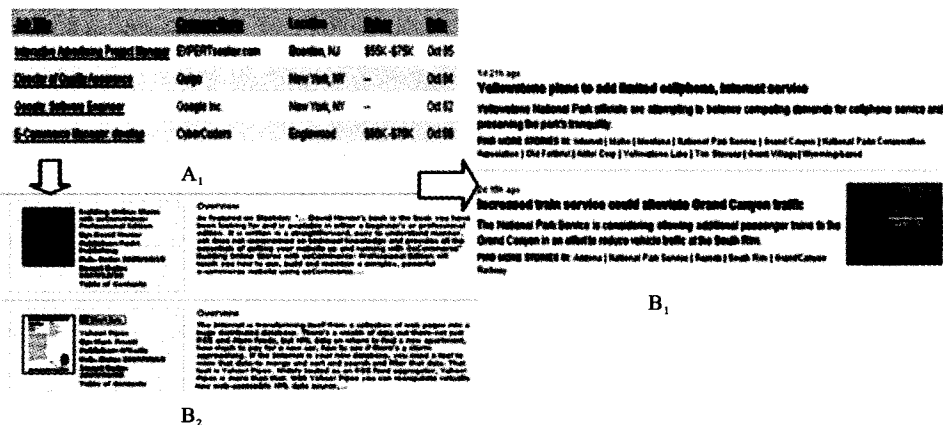


图1 常见 Web 列表页示意图

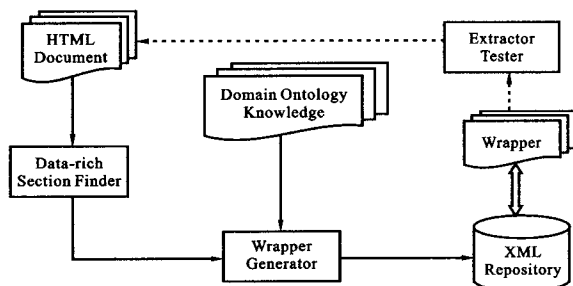


图2 Wrapper 生成过程示意图

对基于模板产生的 Web 页面进行合理的形式化描述,可以更有效地对复杂页面进行数据抽取。

定义1 列表网页模板 T : 列表网页模板 $T = H \cup N$. H 代表用户所关心的抽取数据区域 Data-rich; N 代表噪声区. 用一个二元组 (S, P) 来表示 H , 其中, S 代表数据记录集 (DRs), P 代表 DR 之间的分隔关系。

定义2 数据记录 DR: 数据记录 DR 可以用一个二元组 (D, G) 表示. D 代表记录属性的集合, G 代表属性之间的聚合关系。

文中以电子商务领域中同一主题的 Web 数据库成为研究背景,探索其复杂列表页的 Wrapper 自动化生成问题.对于复杂页面的 Wrapper 自动化生成研究,主要存在以下一些关键步骤:

(1) Data-rich Section 的发现.从数据上看, Data-rich Section 就是 Web 上数据记录的集合.分类列表页面不仅包含数据记录集区域,也包含广告条、导航条等区域.预处理将会过滤掉噪声数据,保留 Data-rich 区。

(2) 聚集块的发现和抽取项的标注.将 Data-rich 中同一字段的抽取项聚集在一起,根据抽取项

的数据模式,样式特征以及标示词,并结合领域本体关系库,进行语义标注。

(3) 构建 Wrapper 的生成规则。

2.2 系统整体框架描述

结合领域的本体知识,对一组给定的基于模板页产生的复杂列表页面进行训练学习,自动化地产生一组适合当前模板的抽取规则,作为模板页面的 Wrapper.图2描述了 Wrapper 自动化产生以及维护的整体框架流程图.系统有三个重要组成部分: Data-rich Section Finder 子模块、Wrapper Generator 子模块、Wrapper 维护子模块^[5-6]。

Data-rich Section Finder 子模块:通过对基于同一模板生成的多个列表页比较,过滤噪声信息,挖掘用户感兴趣的 Data-rich 区.针对 Data-rich 区往往伴随导航提示的特征,根据文献[7]和文献[8]提出的 DSE 算法和 FLCS 算法,提出了 Data-rich Finder 算法来快速定位 Data-rich 区域。

Wrapper Generator 子模块是其中的核心模块,借助领域本体知识,依据相同抽取项在数据模式、样式特征以及标示词等方面的一致性特征,对 Data-rich 中归于同一字段的抽取项进行聚类,提取抽取规则,用配置文件的形式存储在 XML Repository 中。

Extractor Tester 子模块则根据生成的 Wrapper 规则,抽取相应的 Html 文档的数据.如果抽取错误达到某个阈值,则认为该 Wrapper 已经失效,重新生成 Wrapper。

2.3 领域本体知识库的建立

在建立领域本体知识库时,可以利用词与词之

间存在的三种语义关系来帮助我们建立特定领域下的概念关系结构. 这三种关系分别是同义关系、上下位关系、包含关系. 文中以 Yahoo Directory 中的 500 多个 Bookstore 提供的列表页进行训练学习, 构建 Book 概念下的本体关系与标示词库. 以下为用 xml 语言描述的 Book 实体的语义关系片段.

```
<Book>
<Property name="author">
  <tagname>altext</tagname>
  <dataschema><dataschema>
  <descriptionby>by...</descriptionby>
</Property>
<Property name="photo">
  <tagname>img</tagname>
  <dataschema><dataschema>
  <descriptionby></descriptionby>
</Property>
<Property name="ISBN">
  <tagname>text</tagname>
  <dataschema>(0-9)*-(0-9)*-(0-9)*-(0-9)*
<dataschema>
  <descriptionby>isbn...</descriptionby>
</Property>
.....
<Property name="price">
  <tagname>altext</tagname>
  <dataschema>$ (1-9)*.(1-9)*<dataschema>
  <descriptionby>price...</descriptionby>
</Property>
</Book>
```

从中可以看出, Book 概念包含了 author, photo, isbn, price 等属性信息. 每种属性中, 又包含该属性常用的标签名、数据模式以及标示词等信息.

2.4 Data-rich Section 的发现

基于模板生成的列表页中, 不仅包含用户感兴趣的数据记录集, 也包含导航条、广告条等噪声信息. 因此, 需要在数据预处理中过滤掉噪声信息, 提取出数据记录集.

通过大量的实验观察发现, 列表页面的 Data-rich 附近往往具有导航提示区域, 例如分页信息、总记录信息. 利用列表页面的这个特点, 文中对 DSE 算法作了一定程度的修改, 提出了 Data-rich Finder 算法. 该算法主要增加了以导航信息作为修剪分枝的探索项. 具体的算法描述如下:

- 输入: 同一模板产生的两个列表页面 URL
 输出: 代表 Data-rich 区的正向最长公共子串
 算法步骤:
 (1) 输入基于同一模块的两个页面的 Url.

(2) 分别对两个页面的 HtmlTag 树进行深度递归, 如果发现其子树中具有导航提示节点, 则对导航提示节点的祖先节点逆序递归试探, 直到找到可能包含 Data-rich 的节点, 或 HtmlTag 树的根节点.

(3) 以步骤(2)中返回的节点为根节点开始, 对 HtmlTag 树进行深度递归比较, 来判断其子树是否一致. 如果一致, 则丢弃该子树节点.

(4) 将得到的不同子树的正向最长公共子串输出, 作为 Data-rich 的树路径.

2.5 Wrapper 规则配置文件的生成

同一字段的抽取项, 在标签类型、数据模式、特征样式以及特征词方面存在很大的相似性. 通过以下函数来计算文本节点以及图片节点的相似性, 对同一字段的抽取项进行聚类.

$$\begin{cases} \text{Sim}(A, B) = w_1 * \text{SimPtag}(A, B) + w_2 * \text{SimS}(A, B) + w_3 * \text{SimC}(A, B) & \text{文本节点} \\ \text{Sim}(A, B) = w_1 * \text{Simtag}(A, B) + w_2 * \text{SimS}(A, B) & \text{图片节点} \end{cases}$$

其中, w_1, w_2, w_3 分别为相应的权重比例, $\text{SimPtag}(A, B)$ 计算 A, B 节点父节点的标签名是否一致; $\text{Simtag}(A, B)$ 计算 A, B 节点标签名是否一致; $\text{SimS}(A, B)$ 计算 A, B 节点的样式相似度; $\text{SimC}(A, B)$ 计算节点的内容相似性(主要依据标示词与数据类型特征).

依据领域 Ontology 知识信息, 按照特征词、数据模式以及样式特征的权重比例, 制定了如图 3 所示的判决规则链, 对聚集后的文本以及图片信息进行标注.

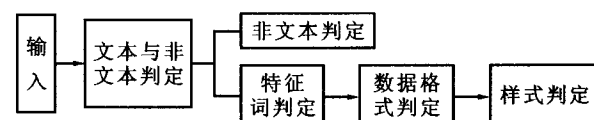


图3 判定规则链

对抽取项进行语义标注之后, 利用其所在结点的标签信息、特征词以及数据类型, 产生该抽取项的抽取规则.

3 实验评价

信息抽取技术采用数据记录的查全率 R 、查准率 P 作为评价标准. 数据记录查全率 R 主要用于测试数据记录的发现程度, 查准率主要用来作为属性抽取的评价标准. cr 表示正确发现的数据记录数, tr 代表总的数据记录数, ce 表示已正确抽取出的属性数, te 表示总抽取的属性数. 可用如下的公式表示:

$$R = \frac{cr}{tr} \times 100\% \quad P = \frac{ce}{te} \times 100\%$$

实验采用 httpclient 和 hmlparser 开源软件对 Web 页面进行下载与解析,对 Safari Books Online 站点的 Books 分类列表页面以及 usatoday 的分类列表页面(每类页面各抽取 50 个)进行数据测试,得到有效性实验的测试结果,如表 1 所示。

表 1 有效性实验结果

所属领域/数据源/类别			查全率 R/%	查准率 P/%
Books	Safari	Java	89.30	83.80
		.Net	92.60	82.63
		E-Commerce	86.70	80.20
News	usatoday	Religion	87.50	85.75
		Education	85.20	83.52
		Word	90.30	79.40

在对 Bookstore Web 数据源的抽取过程中,属性抽取的准确度随着主题数据库下训练集的增多而得到提升,并逐渐趋向稳定,如图 4 所示。

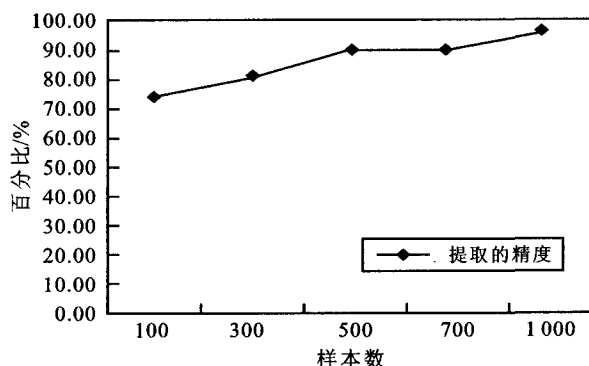


图 4 属性提取精度的变化

4 结束语

文中结合 Ontology 对基于模板的复杂页面的 Wrapper 自动化生成进行了探索性的研究。以主题 Web 数据库的集成为研究背景,以复杂网页为研究对象,依据抽取项的样式特征、标示词进行节点定位,依据数据模式导出数据抽取表达式。最终生成字段抽取规则。实验表明,这种方法能够有效地应用到主题 Web 数据库的集成中去,且抽取的精度随着主题本体知识库完善而逐渐提高。

参考文献:

- [1] Crescenzi V, Mecca G, Merialdo P. Roadrunner: towards automatic data extraction from large Web sites[C]// Proceedings of the 27th International Conference on Very Large Data Bases. San Francisco: Morgan Kaufmann, 2001:109-118.
- [2] Yang Jaeyoung, Tae-Hyung Kim, Joongmin Choi. An interface agent for wrapper-based information extraction[J]. Lecture Notes in Computer Science, 2005(3371): 291-302.
- [3] Meng X F, Hu D D, Li C. Schema guided wrapper maintenance for Web-data extraction[C]// Proceedings of ACM WIDM'2003. New York: ACM Press, 2003: 1-8.
- [4] Arasu A, Garcia-Molina H. Extracting structured data from Web pages[C]// Proceedings of the 2003 ACM SIGMOD international conference on Management of data. New York: ACM Press, 2003:337-348.
- [5] 胡东东,孟小峰.一种基于树结构的 Web 数据自动抽取方法[J].计算机研究与发展,2004,41(10):1607-1613.
- [6] Zhu Jun, Nie Zaiqing, Wen Jirong, et al. Simultaneous record detection and attribute labeling in web data extraction[C]// Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM Press, 2006:494-503.
- [7] Wang J, Lochovsky F. Data-rich section extraction from HTML pages[C]// Proceedings of the 3rd International Conference on Web Information Systems Engineering. Washington: IEEE Computer Society, 2002: 313-322.
- [8] 陈晓锋,张凌,董守斌.基于 XPath 比较的 Web 数据抽取方法[J].郑州大学学报:理学版,2007,39(2):161-166.

作者简介:

陈洪平 男,(1983-),硕士研究生.研究方向为搜索引擎、Web 数据挖掘。

方巍 男,(1975-),博士研究生.研究方向为智能信息处理、Web 数据挖掘、语义 Web、Deep Web。

李林 男,(1982-),硕士研究生.研究方向为网络爬虫、Web 数据挖掘等。

崔志明 男,(1961-),教授,博士生导师.研究方向为智能化信息处理、计算机网络应用。