面向领域的Web数据抽取与集成

李 贵 李征宇 陈韶刚 韩子扬 孙 平 孙焕良
(沈阳建筑大学信息与控制工程学院 沈阳 110168)

摘 要 面向领域的Web数据挖掘包括领域Web数据抽取和领域Web数据集成。针对领域数据抽取,提出了Web结构数据模型和Web表模式,给出了Web表定位和数据记录抽取的算法,针对领域Web数据集成,提出了基于领域模型的数据集成算法。结合行业领域的实际需求,验证了模型和算法的有效性。

关键词 Web结构数据模型,Web表模式,领域模型,领域数据抽取与集成

中图法分类号 TP311 **文献标识码** J

Web Data Extraction and Integration in Domain

LI Gui LI Zheng-yu CHEN Shao-gang HAN Zi-yang SUN Ping SUN Huan-liang
(Faculty of Information & Control Engineering, Shenyang Jianzhu University, Shenyang 110168, China)

Abstract Web data mining in a domain includes Web data extraction and Web data integration. In the phase of Web data extraction, the paper proposes a Web structural data model and a Web table schema, and puts forward the Web table positioning and records extracting algorithm. In the phase of Web data integration, a Web data integration algorithm based on the domain model is presented. The experiment results are given to show effectiveness of the proposed algorithms.

Keywords Web structural data model, Web table schema, Domain model, Domain data extraction and integration

1 引言

随着越来越多的不同行业领域的企业和组织机构在Web上发布大量的信息,从网页中抽取和集成这种数据的能力正变得越来越重要。通过Web数据抽取与集成,使人们能够获取和整合来自成千上万Web数据源的数据,以提供面向领域的增值服务。Web数据抽取任务的研究始于20世纪90年代末,目前针对Web数据抽取的研究主要采取如下3种方法[1-3]:手工方法、包装器归纳和自动抽取。手工方法和包装器归纳的缺点是无法处理大量站点和网页情形,并且如果站点频繁更新的话,维护的开销会很大。自动抽取可能会抽取大量无用数据,需要复杂数据模式和数值的匹配,效率不高。

为了提供基于Web的面向领域的数据增值服务,我们需要从大量异构的网站中提取数据,然后把提取的数据集成为一个统一的数据库,这是因为不同的网站往往使用不同的数据格式。对不同的Web数据表而言,集成意味着匹配出表示同类信息的列,或者匹配语义相同但表达方式不同的值,然而,目前对这方面的信息集成的研究非常少,大部分的研究都是针对Web搜索界面集成问题[4,7]。

本文在分析Web结构化数据特点和领域需求的基础上,引入了Web结构化数据模型、Web表模式和领域模型,然后设计了基于该模型的Web数据抽取和集成算法,并结合实际需求给出算法的实验结果。

2 Web结构数据模型与Web表模式

Web上结构化数据通常是由后台数据库中的数据记录按一定的格式(比如:列表页和详情页)展现在页面上。从数据表现的结构上分析,这类结构化数据通常由数据项、记录元组和记录集合构成。针对Web结构化数据,本文提出一种Web结构数据模型,作为Web结构化数据分析与建模的基础[4-6]。

定义1(Web结构数据模型) 其定义为一个三元组(Type_set, T_tree_set, T_enc_set)。

Type_set为类型集,包括:基类型、元组类型、集合类型、平坦元组类型、平坦集合类型、平坦关系和嵌套关系。

• 基类型 Basic_types=$\{B_1, B_2, \cdots, B_k\}$,其中,$B_i$ 表示原子类型,其值域 $dom(B_i)$ 是一个常量的集合。

• 元组类型 Tuple_type=$[T_1, T_2, \cdots, T_n]$,其中,T_i ($1 \leqslant i \leqslant n$)是基类型或集合类型,$dom([T_1, T_2, \cdots, T_n]) = \{[v_1, v_2, \cdots, v_n] \mid v_i \in dom(T_i)\}$。

• 集合类型 Set_type=$\{T\}$,其中,T 是一个元组类型,其值域 $dom(\{T\})$ 是 $dom(T)$ 的幂集。

• 平坦元组类型:对于一个元组$[T_1, T_2, \cdots, T_n]$,其中 T_i ($1 \leqslant i \leqslant n$)是基类型,那么该元组就是平坦元组类型。

• 平坦集合类型:对于一个元组集合$\{T\}$,T 为平坦元组类型,那么该集合就是平坦集合类型。

本文受国家自然科学基金(61070024)资助。

李 贵(1964—),男,博士,教授,主要研究方向为Web信息挖掘与集成、Web服务与集成,E-mail:ligui21c@sina.com;李征宇(1980—),男,讲师,主要研究方向为Web信息挖掘与集成。

• 平坦关系：对于一个集合，如果该集合的所有元素都是基类型，那么该集合所表示的关系是平坦关系。

• 嵌套关系：对于一个集合，如果该集合的元素中有集合类型的元素(不都是基类型)，那么该集合所表示的关系是嵌套关系。

T_tree_set 为类型树，包括：基类型树、元组类型树和集合类型树。

• 基类型树：一个基类型 B_i 是一棵叶子树或者一个叶节点。

• 元组类型树：一个元组类型$[T_1, T_2, \cdots, T_n]$是一棵以元组节点为根的含 n 颗子树的树，每个 T_i 对应一棵子树。

• 集合类型树：一个集合类型$\{T\}$是一棵以集合节点为根的含一棵子树的树。

T_enc_set 为类型编码规则集，包括：基类型编码规则、元组类型编码规则和集合类型编码规则。

• 基类型编码规则：对于一个用 T 标注的基类型的叶子节点，其实例 t 将被编码成：

T_enc(T;t)=S_TAG t _TAG，S_TAG 和_TAG 分别表示开始标志和结束标志。

• 元组类型编码规则：对于一个用 T 标注的有 N 个属性的元组节点$[T_1, T_2, \cdots, T_n]$，这个元组类型的实例将被编码成：

T_enc(T;$[t_1, t_2, \cdots, t_n]$)=S_TAG T_enc(t_1)，T_enc(t_2)，…，T_enc(t_n) E_TAG，S_TAG 和 E_TAG 分别表示开始标志和结束标志。

• 集合类型编码规则：对于一个用 T 标注的集合类型的节点，这个非空的集合实例$\{s_1, s_2, \cdots, s_n\}$将被编码成：

T_enc(T;$\{s_1, s_2, \cdots\cdots, s_n\}$) =S_TAG T_enc(s_1)，T_enc(s_2)，…，T_enc(s_n) E_TAG，S_TAG 和 E_TAG 分别表示开始标志和结束标志。集合元素被一个排序函数排列起来。

列表：按照一个排序函数<排好序的集合实例，一个空集合实例将被编码成 S_TAG E_TAG。

Web 表数据被建模成平坦关系和嵌套关系，可以包含集合和元组的有类型的对象。平坦关系和嵌套关系通常可以用 DOM 树来表示，关系实例的 HTML(或 XML)标记编码将 DOM 树的每个节点与一个基于编码规则的标注函数相关联。

后台数据库中的数据都是以结构化的形式展示于 Web 表中，Web 表包括列表类型和详情类型，可以以 Web 表作为基本的数据抽取对象，为此在 Web 数据模型的基础上定义 Web 表数据模式。

定义 2(Web 表模式) 其为一个五元组(W_Table_T，W_Table_N，W_Table_DN，W_Table_R_set，W_Table_enc)：

W_Table_T：Web 表类型(详情页或列表页)。

W_Table_N：Web 表名，是一个集合类型，由表示某一领域同一实体的不同 Web 页面中若干 Web 表名构成。

W_Table_DN：Web 数据项名，是一个集合类型$\{B_1, \cdots, B_n\}$，其中，B_i 是同一 Web 数据项的同义词的集合。

W_Table_R_set：Web 表数据记录、集合，是一个元组集合类型，每条记录是一个元组类型。

W_Table_enc 为 Web 表数据记录的 HTML 或 XML 编码规则，是 Web 结构数据模型中类型编码规则集 T_enc_set 的实例。

3 领域模型

Web 页面表数据通常是由查询相关的后台数据库表记录、通过 HTML 或 XML 编码规则产生的，所以面向领域的 Web 结构数据具有以下的特性：

(1)有限的实体和属性。在一个领域中一般只有有限的实体名(Web 表名)和属性名(数据项名)。(2)面向领域的 Web 网站中存在大量相似的 Web 表格实体和同义词数据项名。比如行业网站提供相似的数据查询服务或者为同类的产品提供了大量相似的表格。(3)附加的结构语义。表格上的属性往往是有语义约束的，属性之间存在语义约束或层次化的结构关系[6]。

领域数据集成需要寻找各个表格属性之间的映射，而寻找属性之间的映射中最基本的问题是寻找同义词。一旦对同一个领域的一组 Web 表格完成匹配，就可以自动地构建一个领域数据模型。领域中数据的组织和表达基于领域属性、实体和语义约束，由领域属性、实体、约束以及与领域中不同站点或页面的 Web 表的映射关系构成领域模型。

定义 3(领域模型) 其为一个三元组(D_Attr_set，D_Tab_set，D_mapping_set)，其中：

• D_Attr_set 为领域属性集：表示为一个四元组(A_i_N，A_i_Type，A_i_mapping，A_i_constr)的集合$\{(A_1$_N，A_1_Type，A_1_mapping，A_1_constr)…(A_n_N，A_n_Type，A_n_mapping，A_n_constr)$\}$。

其中，A_i_N 为属性名，A_i_Type 为属性类型，A_i_mapping 为 Web 表数据项名与该领域属性的对应关系，A_i_constr 为属性约束。

• D_Ent_set 为领域实体集：表示为$\{$Entity_T_1，…，Entity_$T_n\}$，Entity_T_i=(A_i_SET，T_i_constr)，其中 $1 \leqslant j \leqslant n$，$A_i$_SET 为实体 T_i 的属性集合，T_i_constr 表示 T_i 的属性值之间的约束，且任意两个实体 Entity_T_i 与 Entity_T_j 之间除外来关键字外不允许有重复的属性。

• D_mapping_set 为领域映射集：表示为$\{$(Entity_T_i，$\{$ W_Table_$N_i\}$)(A_N_i，$\{$W_Table_DN_{i1}，…，W_Table_$DN_{ir}\}$)$\}$，$1 \leqslant i \leqslant n$，$1 \leqslant r \leqslant n$，表示领域实体名与 Web 表名和实体属性与 Web 数据项的映射关系。

4 Web 结构化数据抽取与集成算法

基于上面提出的 Web 表模式和领域模式，Web 数据抽取与集成的过程分为 3 个阶段：Web 表定位、表数据抽取和领域数据集成。

4.1 Web 表定位算法

依据 Web 表名的定位方式：

```
输入：URL，W_Table_N
输出：table_body
算法：
(1)URL_HTML=get_HTML(URL[i]);
(2)  local_head=NULL;
(3)  while i<=n，and local_head=NULL
     do{ i=i+1 ;
     local_head=get_loca(URL_HTML，W_Table_Ni);}
(4) table_body=
     get_body(URL_HTML local_head);
```

给定网址 URL，第(1)步获取 html 源码并将其解析处理成 XML 格式；第(3)步搜索源码，以获得表头位置；第(4)步获取 local_head 为表头的表体。

4.2 Web 表记录抽取算法

Web 结构化数据主要体现在列表页和详情页两种形式中，为此提出基于一般列表和详情表两种数据的抽取算法。

4.2.1 一般列表的数据抽取

输入：一个 Web 表体 table_body，该表体为已编码的 N 个集合类型的实例，其中一个实例表示该 Web 表数据项名称，其余实例是 Web 表数据记录实例。

输出：W_Table_DN_set，W_Table_R_set

输入字符串集 table_body 是一个 Web 表，由元组类型的 W_Table_DN_set 与 W_Table_R_1，W_Table_R_i，W_Table_R_n 按照某种标记编码函数 T_enc(W_Table_DN_set)，T_enc(W_Table_R_1)，T_enc(W_Table_R_i)，T_enc(W_Table_R_n)构成。

算法：

(1)data_table=get_data_table(table_body)

(2)W_Table_DN_set=get_W_Table_DN(data_table)

(3)W_Table_R_set= get_W_Table_R(data_table)

算法第(1)步获取数据表格；第(2)步获取数据项名；第(3)步获取数据项记录集。

4.2.2 详情表的数据抽取

输入：一个 Web 表体 table_body 的字符串集 S，S 为已编码的一个元组类型的实例(其中包含数据项名称)。

输出：W_Table_DN_set，W_Table_R

详情表的数据项名，数据项记录通常都在同一个数据表格，可以直接进行抽取操作。详情表是由一个元组类型 t 的一个实例按照某个标记编码函数 T_enc(W_Table_R_set)组成的。

算法：

(1)W_Table_R_set=get_W_Table_R_set(table_body)；

(2)W_Table_DN_set=get_W_Table_DN(W_Table_R_set)；

(3)W_Table_R=get_W_Table_R(W_Table_R_set)；

算法第(1)步获取数据项记录集；第(2)步在数据项记录集中获取数据项名；第(3)步在数据项记录集中获取数据项值；

4.3 领域数据集成

领域数据集成主要包括领域数据整合和领域模型完善两方面的问题。

领域数据整合是将已经抽取的数据项值 W_Table_R_set，依据 D_Ent_set 中的 T_i_constr 和 D_mapping_set 经过类型转换和一致性检查等工作之后加入到依据领域模型定义的领域数据库表里。

领域模型完善是针对领域 Web 站点或页面的动态变化(包括原有 Web 表模式的改变和新 Web 表的增加)，对领域模型进行更新和增加新的领域属性、实体和约束的工作。

输入：W_Table_R_set，W_Table_DN_set，D_mapping_set

算法：

(1)Connect to database

(2)While(length(W_Table_R_set)>0)

(3)For i=1 to length(W_Table_R_set) do add_W_Table_R_set(Entity_T_i,{A_N_{i1},A_N_{ij}})；

(4)For i=1 to length(W_Table_R_set) do{ if W_Table_N then

update_D_mapping_set(D_mapping_set,

W_Table_N,W_Table_DN_set);

else

add_D_mapping_set(D_mapping_set,

W_Table_N,W_Table_DN_set);}

算法第(1)步连接数据库；第(3)步将 W_Table_R_set 中的数据导入到数据库中；第(4)步导入完成之后，如果 W_Table_R_set 中还有未导入的数据，那么判断是否存在该表名，如果存在，则更新领域数据模型；如果不存在，则添加这个新的领域数据模型。循环这个过程直到 W_Table_R_set 中所有的数据都被导入数据库中。

5 试验结果

为了验证基于 Web 表模式的数据抽取和基于领域模型的数据集成算法的有效性，本文结合实际应用项目“中国地产信息平台及综合应用系统”的需求，对全国不同城市的房地产网站进行了抓取实验，本实验结果是采用 3 台机器在 100M 共享网络带宽环境下分布式并行抓取的表现。

实验中对某城市房产网站中的许可信息网页(详情页类型)(网址：www. syfc. com. cn/work/ysxk/ query_ xukezheng. jsp ? cur_page=1)、楼盘信息网页(列表页类型)(网址：www. syfc. com. cn/work/xjlp/new _building. jsp)、楼栋信息网页(列表页类型)(网址：www. syfc. com. cn/work/xjlp/build_li -st. jsp)、户型信息网页(详情页类型)(网址：www. syfc. com. cn/work/xjlp/ * / *)进行了抓取实验。

表 1 给出了抓取的结果，“Num”表示抓取到的记录条数，“Time(s)”表示抓取所耗费的时间，“Avg_time(s)”表示平均耗费时间。从表中可以看出，网速的提高会很明显地改善程序的效率，同时数据量的大小也对实验有较大影响。

表 1 数据抽取的性能

	许可信息	楼盘	楼栋	户型
Num	4859	4323	30553	750862
Time	186	175	1326	3 * 3590
Avg_time	0.038	0.04	0.043	0.014

为了检测本文算法的精准度和完整度，对不同的城市进行了抓取实验。本文采用以下几种测量性能的公式：准确率(P)、召回率(R)和 F-score(F)。F-score 是准确率和召回率的调和平均值。

表 2 中给出了实验的结果。其中“Number of Record”指抓取的记录数，“Records Proposed”指正确的记录数，“Correct”指抓取正确的记录数。

表 2 不同城市的抽取测试

城市	鞍山	长沙	大连	东莞	佛山
Num	183415	665640	1120721	507758	417616
Records Proposed	174565	659858	1113460	506024	399253
Correct	171654	657156	1108456	498546	391148
Precision	0.98	0.996	0.996	0.985	0.98
Recall	0.936	0.987	0.989	0.982	0.937
F-score	0.95	0.991	0.993	0.984	0.958

从表 2 中可以看出，利用本文的方法对领域网站进行抽取无论是准确率还是召回率都有很好的表现。同时，对抓取不正确的结果分析发现有些 Web 表格的表名是以图片形式存在的，针对 Web 图表模式识别与数据提取需要进一步的研究。

结束语 本文通过对现有的 Web 结构化数据抽取方法

(下转第 175 页)

图5　传统形态学处理结果

图6　本文算法结果

文献[1]提出的动态自适应权重多结构多尺度形态学边缘检测方法，从图8结果图看，可得到清晰、光滑的边缘效果。对比图8、图9，本文算法处理的边缘效果更加提高，杂质明显减少。本文采用同一尺寸的4个结构不同的结构元素，计算腐蚀过程相比本文算法速度明显更快。文献[1]在做自适应确定的权重值时用到可填入次数 A，并求倒数运算，其中算法时间如表1所列。

表1

	canny 算子	文献[1]算子	本文算子
用时：	0.2842	11.509	0.3100

图7　文献[1]提供的图

图8　文献[1]结果图

可见文献[1]算法消耗运算量非常大，而本文并没有使用此运算方法，本文算法相比文献[1]速度明显增快，并取得了相比其更加良好的效果。通过眼前节OCT图像的实验表明，新算法比文献[1]算法在算法时间上有了较大提高，边缘细节上也更完整，提高了边缘检测的正确性，很好地保留了图像的边缘图像，为之后的检测奠定了良好的基础。

图9　本文方法处理图7结果

结束语　本文提出此方法是为了更好更快速地识别OCT图像的边缘特性。本文采取改进的形态学方法，即采用多结构元素，通过对图像的预处理来提高图像暗区域的细节，减少图片本身的细节不明显问题。再通过基本的腐蚀与膨胀，提高了图像的分辨度并且去除了噪音，提高了边缘提取的精确度。

实验结果表明本文方法优于传统的边缘检测方法，能检测出更多的边缘细节，处理速度快。结合多结构元素的特性，使用合适的边缘检测算子，可以较好地解决边缘检测精度与抗噪声性能的协调问题，从而得到更好的边缘检测结果。相比文献[1]，算法的时间显著提升，对于边缘效果有一定提高且有效减少了图像杂质对图像分析的干扰。

参考文献

[1] Huang Si-wei, Zhang Ang, Tian Xiao-lin, et al. Dynamic Adaptive Weight Multi-scale and Multi-structure Morphological Edge Detection in Anterior Chamber OCT Images Advanced Materials Research[J]. Trans. Tech. Publications, Switzerland, 2012 (340):70-75

[2] 鲁昌华，刘玉娜. 基于同态滤波和改进形态学的图像边缘检测[J]. 仪器仪表学报，2011，32(6)

[3] 高玮玮，沈建新，王玉亮. 基于数学形态学的快速糖尿病视网膜病变自动检测方法[J]. 光谱学与光谱分析，2012

[4] Yu Zi-yi. Analysis of mathematical morphological algorithm for edge detection in micrograph[J]. Computer Technology and Development, 2006, 16(2):100-102

[5] 何新英，王家忠，孙晨霞，等. 基于数学形态学和 Canny 算子的边缘提取方法[J]. 计算机应用，2008，28(2)

[6] 熊立志，陈立潮，潘理虎，等. 基于多尺度轮廓结构元素的多形状边缘检测[J]. 计算机应用研究，2012，29(9)

[7] 林玉池，崔彦平，黄银国. 复杂背景下边缘提取与目标识别方法研究[J]. 光学精密工程，2006，14(3)

(上接第159页)

的研究，结合领域中的 Web 数据表结构的分析结果，从面向领域的数据抽取和集成的角度，提出了基于 Web 表模式和领域模型的数据抽取和集成算法。实验结果表明，本文提出的数据抽取和集成方法提高了领域中数据抽取与集成的效率，能够满足房地产等领域的实际应用需求。同时如何对领域数据模型中实体和属性进行有效的扩充，以及如何有效地实现面向领域的大规模 Web 数据抽取与集成将是进一步研究的方向。

参考文献

[1] Cafarella M J, Halevy A, Wang D Z, et al. WebTables: Exploring the Power of Tables on the Web[C]// Proceedings of VLDB-08. Auckland, New Zealand, 2008:538-549

[2] Crestan E, Pantel P. Web-Scale Knowledge Extraction from Semi-Structured Tables[C]// Proceedings of WWW-2010. Raleigh, North Carolina, USA, 2010

[3] Liu Bing. Web Data Mining[M]. 俞勇，薛贵荣，韩定一，译. 北京：清华大学出版社，2009：265-26

[4] Chen H, Tsai S, Tsai J. Mining Tables from Large-Scale HTML Texts[C]// Proceedings of COLING-00. Saarbrücken, Germany, 2000

[5] Robert G, Wilks Y. Information extraction: Beyond document retrieval[J]. Journal of Documentation, 1998, 54(1):70-105

[6] Gatterbauer W, Bohunsky P, Herzog M, et al. Towards Domain-Independent Information Extraction from Web Tables[C]// Proceedings of WWW-07. Banff, Canda, 2007:71-80