Computer Science and Application 计算机科学与应用, 2016, 6(4), 203-215
Published Online April 2016 in Hans. http://www.hanspub.org/journal/csa
http://dx.doi.org/10.12677/csa.2016.64026

Hans汉斯

The Research and Implementation of Extraction and Integration of Web Data Based on Domain Pattern

Gui Li, Chuanjie Geng, Ziyang Han, Zhengyu Li

Faculty of Information & Control Engineering, Shenyang Jianzhu University, Shenyang Liaoning
Email: ligui21c@sina.com

Received: Mar. 29th, 2016; accepted: Apr. 18th, 2016; published: Apr. 22nd, 2016

Abstract

One of the objectives of the Web data mining is to provide the domain-oriented information value added service. Domain-oriented web data extraction and integration is the basis of providing value added services, and is also a major research direction in the field of web data mining. In combination with the requirement of the field, we proposed the domain-oriented web data extraction and integration architecture. Based on the concepts of web data model and web data pattern, domain data model and domain data pattern, the mapping method of web data pattern and domain data pattern and integration method on data level are proposed to solve the conflict problem of pattern layer and data layer in the integration process. We also discussed the implementation method of web data extraction and domain value added services. Real estate information platform and integrated application system are developed with the actual requirements, and the effectiveness of the model and algorithm is verified.

Keywords

Web Data Model and Pattern, Domain Data Model and Pattern, Domain Data Extraction and Integration, Domain Value Added Service

基于领域模式的Web数据抽取与集成系统研究与实现

李　贵，耿传杰，韩子扬，李征宇

文章引用: 李贵, 耿传杰, 韩子扬, 李征宇. 基于领域模式的 Web 数据抽取与集成系统研究与实现[J]. 计算机科学与应用, 2016, 6(4): 203-215. http://dx.doi.org/10.12677/csa.2016.64026

沈阳建筑大学信息与控制工程学院，辽宁 沈阳
Email: ligui21c@sina.com

收稿日期：2016年3月29日；录用日期：2016年4月18日；发布日期：2016年4月22日

摘 要

提供面向领域的信息增值服务是Web数据挖掘的目标之一，面向领域的Web数据抽取与集成是提供领域信息增值服务的基础，也是Web数据挖掘领域的一个主要研究方向，结合领域需求，本文提出一种面向领域的Web数据抽取与集成架构，在给出Web数据模型与Web数据模式、领域数据模型和领域数据模式等相关概念基础上，提出Web数据模式与领域数据模式的映射方法和数据层次上的集成方法，用于解决集成过程中的模式层次和数据层次的冲突问题，并讨论了web数据抽取和领域增值服务的实现方法。结合实际需求开发了房地产信息平台及综合应用系统，验证了模型和算法的有效性。

关键词

Web数据模型与模式，领域数据模型与模式，领域数据抽取与集成，领域增值服务

1. 引言

大数据时代，随着越来越多的企业和组织机构在 Web 上发布大量的信息，从成千上万的网页中抽取和集成这种海量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不同 Web 数据源的数据，以提供面向领域的增值服务。

为了提供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各种数据并集成为一个统一的数据库平台，并在此基础上开发面向领域的增值服务，这其中面临的主要问题包括[1]-[7]：如何从海量异构网页种高效抽取各种源数据；如何解决不同网页抽取数据在模式层次和实例层次的冲突问题；如何建立的领域数据集成模，以及集成模式与各种不同类型网页数据模式的映射关系；如何在领域数据集成模式的基础上建立面向领域的服务模型等主要问题。

目前，针对 web 数据抽取与集成问题的相关技术研究主要体现在如下两方面。

在 web 数据抽取研究方面的研究[1] [2] [7] [9] [10]，Web 数据抽取是一个关于从网页中抽取目标信息的问题。其中包含两大问题：即从自然语言文本中抽取信息和从网页中抽取结构化数据。这里重点讨论结构化数据抽取。Web 结构化数据通常是从后台数据库获取数据记录，并按照一定的模板展现在网页上。Web 数据抽取技术是从 20 世纪 90 年代开始研究，目前有关 web 数据抽取技术可以分为三种主要类别：1) 包装器编程语言和可视化平台，2) 包装器归纳，3) 自动抽取。其中前两种方法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。

同时，随着 web 技术的发展，在 web 数据抽取领域又出现一些新的问题亟待解决，其中包括：隐含 web 数据库抽取、基于动态链接的 web 页面数据抽取、基于用户验证码的 web 数据抽取和基于图模式的 web 页面数据抽取等问题。

在 web 数据集成研究方面的研究[2]-[4] [6]，对于领域应用需求来讲，需要从大量的网站中提取数据并集成，以便通过领域增值服务，在此情况下，需要把从各个网站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

信息的表或列，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 Web 收索界面集成问题。

目前，web 数据集成研究所面临的挑战是如何解决 web 数据的动态集成问题，这不仅涉及不断涌现的新的 web 数据源的，也涉及同一个 web 数据源的信息的内容虽时间也在不断的丰富和变化。

本文在分析 Web 结构化数据特点和领域需求的基础上，提出了基于领域模型的 web 数据抽取和集成架构；在引入了 Web 结构化数据模型、Web 表模式和领域模型的基础上，给出了 web 数据抽取与集成、模式建立与模式映射的算法，并结合房地产领域的实际应用需求设计实现了面向房地产领域的 web 数据抽取与集成系统，验证了上述模型和算法的有效性。

2. 面向领域的 web 数据抽取与集成架构

面向领域的 web 数据抽取与集成架构如图 1 所示，在该架构中所要完成的工作包括，

1) 集成模式与 web 模式的建立：依据某一领域的网站页面数据的特点和该领域的数据特点，在关系数据模型的基础上，建立领域 web 数据模型和领域数据模型，并依据上述模型针对领域中不同的网站分别建立 web 数据模式和统一的领域数据集成模式；

2) Web 数据抽取及抽取方法库的建立：由于领域数据源包含大量的采用不同开发技术和网页数据展现方式的网站，很难采用一种通用的抓取技术来实现对这些网站数据的高效抓取，因此，必须对不同类型的网站采用不同的数据抓取方法，建立面向领域的 Web 数据抽取方法库。依据网站 web 数据模式和网站采用的技术特点，选择合适的 web 数据抽取方法，实现网站 web 数据高效抽取，并采用 XML 等规范化的标准格式进行存储；

3) 模式映射及映射规则库的建立：由于从不同网站抓取的数据在模式上和领域数据集成模式存在模式冲突问题，所以需要对每一个网站的 Web 数据模式建立到集成模式的模式映射规则集，以解决数据集成过程中的模式冲突问题，其中包括，实体冲突、属性冲突和完整性约束冲突等问题；

4) 数据集成及集成规则库的建立：由于从不同网站抓取的数据在属性值和属性约束上和领域集成数据库存在实例冲突问题，所以需要对每一个网站的 Web 抓取数据建立到集成数据库的数据清洗规则集，以解决数据集成过程中的实例冲突问题。

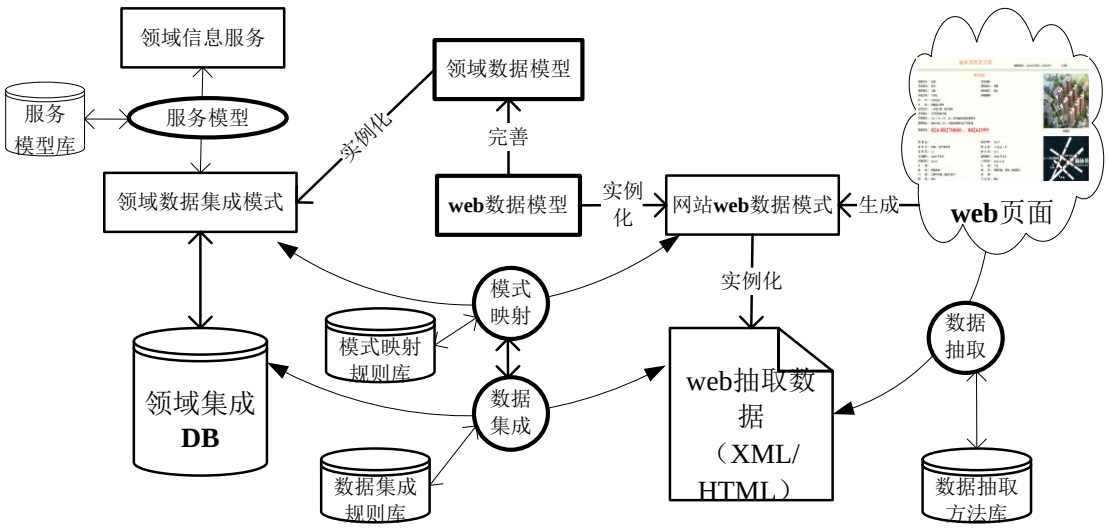

Figure 1. Web data extraction and integration in domain
图 1. 领域 web 数据抽取与集成结构

5) 领域服务及服务模型库的建立：领域数据抽取与集成的目的是提供面向领域的增值服务，所以必须针对领域中不同客户群体的需求，建立不同信息服务模型，实现面向不同客户群体的领域决策支持、公众信息查询和内容推荐等服务，实现领域信息的增值服务。

由于 web 信息的动态性，即 web 信息不断丰富和演变的特点，领域 web 数据抽取和集成系统应具有如下两方面自适应特性：

1) 领域集成模式的演进性，一方面随着市场和企业的经营的发展，企业数据资源的内容和种类也在不断的丰富和增长；另一方面随着 web 技术的发展，企业和政府部门用于组织和展现数据的技术方法也在不断的更新，这就要求系统的集成模式必须采用一种演进的方式来适应这种变化。

2) 领域数据抽取与集成的演进性，领域数据抽取与集成是一种对领域数据不断积累的过程，集成系统必须采用数据的增量抽取与集成方式来解决领域新增数据的集成问题。

3. 数据模型与数据模式

3.1. Web 数据模型与 Web 数据模式

Web 上结构化数据通常是由后台数据库中的数据记录按着一定的格式(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[2] [5] [6]。

定义 1. WEB 结构数据模型：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 = {B1, B2, ……, Bk}，其中，Bi 表示原子类型，其值域 dom(Bi)是一个常量的集合。
- 元组类型 Tuple_type = [T1, T2, ……, Tn],其中，Ti ($1 \le i \le n$)是基类型或集合类型，dom([T1, T2, ……, Tn])={[v1, v2, ……, vn] | vi∈dom(Ti) }；
- 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[T1, T2, ……, Tn]，其中 Ti ($1 \le i \le n$)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。
- 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 Bi 是一颗叶子树或者一个叶节点。
- 元组类型树：一个元组类型[T1, T2, ……, Tn]是一颗以元组节点为根的含 n 颗子树的树，每个 Ti 对应一颗子树。
- 集合类型树：一个集合类型{T}是一颗以集合节点为根的含一颗子树的树。
- 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 E_TAG，S_TAG 和 E_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 个属性元组节点[T1, T2, ……, Tn]，这个元组类型的实例将被编码成：

T_enc(T;[t1, t2, ……, tn])=S_TAG T_enc (t1)，T_enc (t2)，……，T_enc (tn) 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{s1，s2，……,sn}将被编码成：
- T_enc (T; {s1, s2, ……, sn}) =S_TAG T_enc (s1)，T_enc (s2)，……T_enc (sn) 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 表包括详情页和列表页两种类型，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：为一个六元组(W_Table_T, W_Table_N, W_Table_DN, W_Table_R_set, W_Table_enc, W_Mapping _set,)：

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{B1,……Bn}，其中，Bi 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。

W_Mapping _set=

(Attr_Mapping_set, Table_mapping_set)

Web 表模式到领域集成模式中的属性映射和表映射规则集合，其中

Attr_Mapping_set=

{(W_Table_DNiAj_N　Integrity_Rule_set), …..}

IntegrityRule_set 为完整性处理规则集，包括：类型转换处理规则、异常值处理规则、空值处理规则，等等；

Table_mapping_set=

{(W_Table_Ni > Tj_N, …..}为 web 模式中的 web 表 W_Table_Ni 到集成模式中的实体 Tj_N 的映射规则。

3.2. 领域模型与领域数据模式

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，属性之间存在语义约束或层次化的结构关系[6]。

领域数据集成首先是要建立领域集成模式，并完成 Web 表格到集成模式的映射与匹配。一旦对同一个领域的一组 Web 表格完成匹配，就可以自动的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：为一个三元组(D_Attr_set, D_Tab_set, D_constr_set)

其中：

- D_Attr_set 为领域属性集：表示为一个三元组(Ai_N, Ai_Type, Ai_constr)的集合{(A1_N, A1_Type, A1_constr)……(An_N, An_Type, An_constr)}.
- 其中，Ai_N 为属性名，Ai_Type 为属性类型，Ai_constr 为属性完整性约束(包括属性的值域、初始值等)；
- D_Ent_set 为领域实体集：表示为三元组 Entity_Ti = (Ti_N, Ai_SET, Ti_constr) 的集合{Entity_T1,……Entity_Tn}，其中 1≤j≤n，Ti_N 为实体名，Ai_SET,为属性集合，Ti _constr 表示属性值之间的依赖关系和关键字属性等，且任意两个实体 Entity_Ti 与 Entity_Tj 之间不允许有重复的非键属性。
- D_constr_set 为领域约束集：表示领域中不同实体 Entity_Ti，和 Tj 之间($1 \le i \le n, 1 \le j \le n, i \ne j$)实例的对应关系(包括 1:1，1:n 和 n:m)。

定义 4. 领域模式：是针对某一具体领域集成需求利用领域模型概念建立的集成数据库模式，表示为一个四元组(M_N, M_Attr_set, M_ Ent_set, M_constr_set)。

其中：

- M_N 为领域模式名(即，领域数据库名)。
- M_Attr_set 为某一个具体领域属性集；
- M_Ent_set 为某一个具体领域的实体集，
- M_constr_set 为某一个具体领域约束实例集。

4. 模式映射与数据集成

4.1. 模式建立与模式映射

为了实现面向领域的 web 数据集成，首先要依据领域需求和领域所涉及的 web 数据源的页面特点，分别建立领域集成模式和 web 表模式，并完成 web 表模式到领域集成模式的映射[10]-[13]。领域模式建立与模式映射的过程如图 2 所示。

1) 模式建立，依据网站页面的数据格式和领域的数据需求，建立网站页面对应的 web 表模式和领域集成模式；

2) 模式映射，建立 web 模式到领域集成模式的模式映射规则，在建立模式映射规则的过程中首先完成属性匹配，然后完成实体匹配，并逐步完善领域的集成模式；当网站页面的 web 表模式发生变化时需要重新进行模式映射。

属性匹配是完成 web 模式中的数据项到集成模式中的属性匹配，而实体匹配是完成 web 表模式中的 web 表到集成模式中的实体匹配。这其中涉及到的主要操作包括：集成模式中的属性扩充、实体扩展、领域约束完善和 web 表模式与集成模式映射规则的完善。上述操作的具体算法思路描述如下。

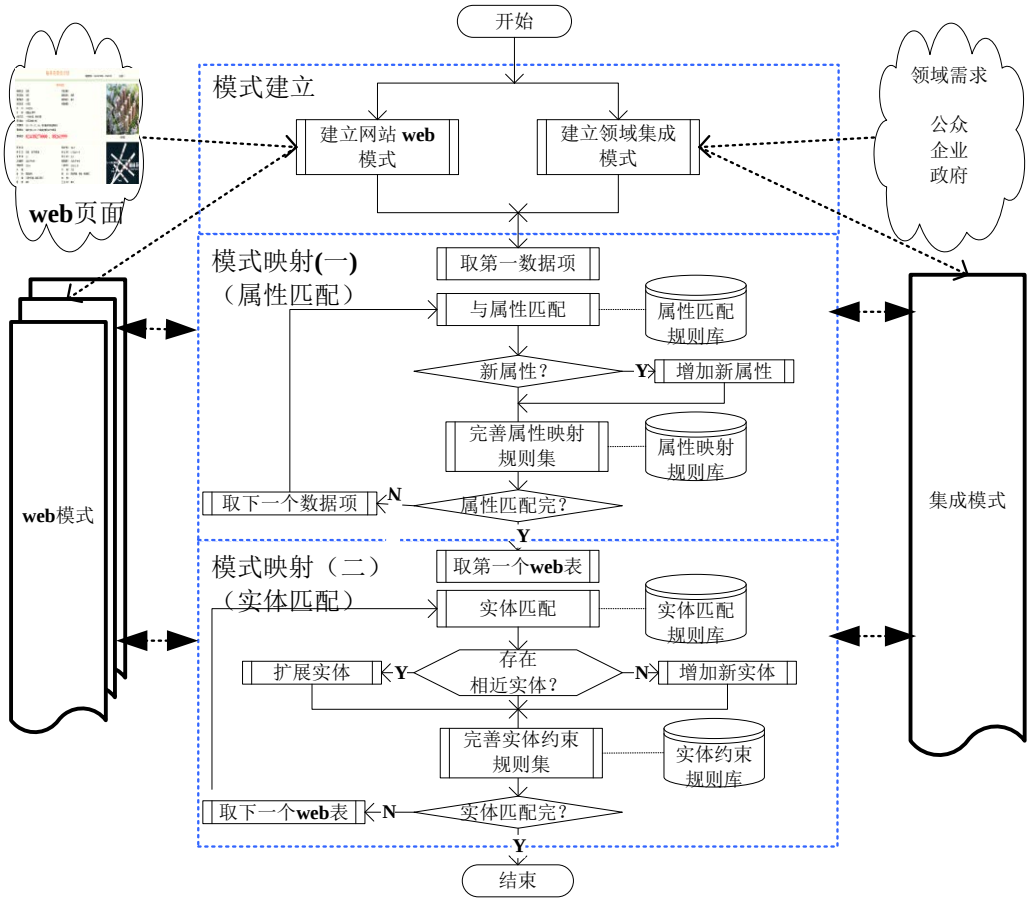

Figure 2. Process of pattern creation and mapping
图 2. 模式建立与模式映射流程图

属性扩充：扩充领域实体的属性，前提条件是 Web 表模式中含有领域模式中不包含的属性，通过对领域模式中的实体扩展属性可解决；

算法思路

1) 在领域模式 M_N 中增加新属性 Ai_N

ADD_M_Attr_set(Ai_N, Ai_Type, Ai_constr)

2) 在对应的领域实体 M_Ej 中增加属性 Ai_N

APPEND_ M_Ent_set (M_Ej A_Nj)；

3) 修改 Web 表模式中的映射规则集

MODI_W_ Table(W_Mapping _set)；

4) 完善领域约束集 M_constr_set

MODI_ M_constr(M_constr_set)；

实体扩展：增加领域模式实体，前提条件是 Web 表模式中含有领域模式中不包含的实体，而且不能通过实体属性扩展来解决，为此要通过增加领域模式中的实体来解决。

算法思路

1) 在领域模式 M_N 中逐个增加新出现的属性

ADD_M_Attr_set(Ai_N, Ai_Type, Ai_constr)

2) 在领域模式中增加新实体

ADD_ M_Ent_set (Tj_N, Aj_SET, Tj_constr))

(Ai_N∈Aj_SET)；

3) 修改 Web 表模式中的映射规则集

MODI_W_ Table(W_Mapping _set)；

领域约束完善：由于集成模式中原有实体属性的增加和新实体的扩展，需要对领域约束进行补充和完善，具体包含两种情况：

1) 在原有实体中增加属性

当增加的属性为关键字(或关键字的一部分)时，需要调整实体之间的对应关系；

当增加的属性为非关键字时，不需要调整实体之间的对应关系；

2) 增加新的实体

需要根据新增实体的关键字属性与其它实体中的关键字属性的依赖关系确定新增实体与其它实体的对应关系。

web 模式映射规则完善：在模式映射过程中，只要发生属性增加和实体扩展，都需要修改和完善 web 模式中的映射规则。分两种情况处理：

1) 增加新的属性时，增加属性映射规则；

2) 增加新的实体，增加实体映射规则。

4.2. 数据集成

在完成模式层次上的映射之后所面临的问题是如何把抽取到的数据集成到领域集成数据库里，以便利用这些集成后的数据提供领域增值服务。

数据层次上的集过程如图 3 所示，所要解决的关键问题包括[13] [14]，1) 如何高效查找集成数据库中与抽取到的数据的记录相匹配的记录(比如抽取数据以 XML 文档格式存储)；2) 当存在匹配记录时，如何依据属性的完整性规则修改集成数据库中匹配记录的相应属性值；3) 当不存在匹配记录，需要插入新的记录时，如何依据属性的完整性规则建立新记录的相应属性值。在修改已存在记录的属性值和设置新记录的属性值时，都依据属性的完整性约束解决属性的类型转换、异常值处理、空值处理、值域转换等问题。为了解决上述问题，需要依据属性的完整性约束和实体中属性的依赖关系建立领域数据清洗规则，其中包括，匹配记录的判定规则、属性完整性约束规则等。

1) 相同记录的判定规则：为了快速实现 web 抽取纪录与集成库中的相同纪录的判定，需要采用合理的数据库元组分块策略，以减少相同纪录的判定的比较代价(比较次数)；

2) 完整性约束规则：包括：

类型转换规则：将抽取数据的数据项类型转成集成数据库对应实体的属性类型；

异常值的处理规则：当抽取的数据出现异常值时，如何依据完整性规则进行修正；

空值的处理规则：当抽取的数据出现异常值时，采用属性依赖规则或属性约束进行适当处理，

属性值域扩展规则：当属性出现新的合理值时，需要扩充相应属性的值域，即修改完善属性完整性规则。

5. WEB 数据抽取与领域增值服务

5.1. WEB 数据抽取

Web 数据抽取过程如图 4 所示，分三个主要阶段：

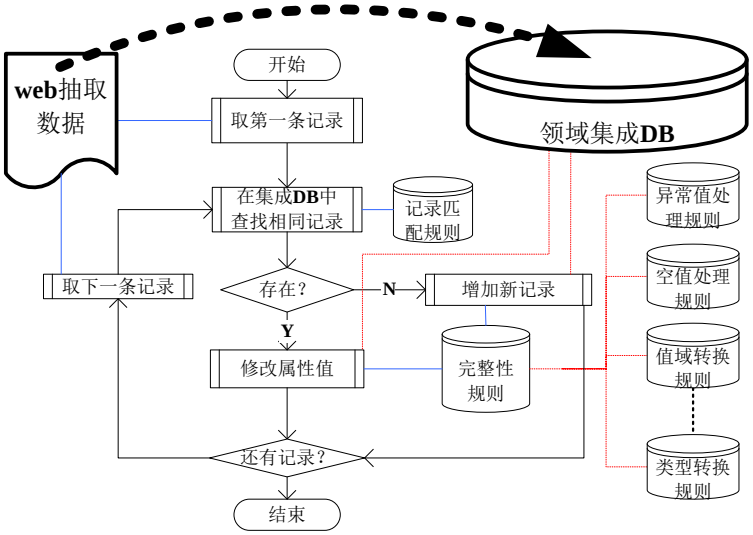

Figure 3. Process of data integrating
图 3. 数据集成流程图

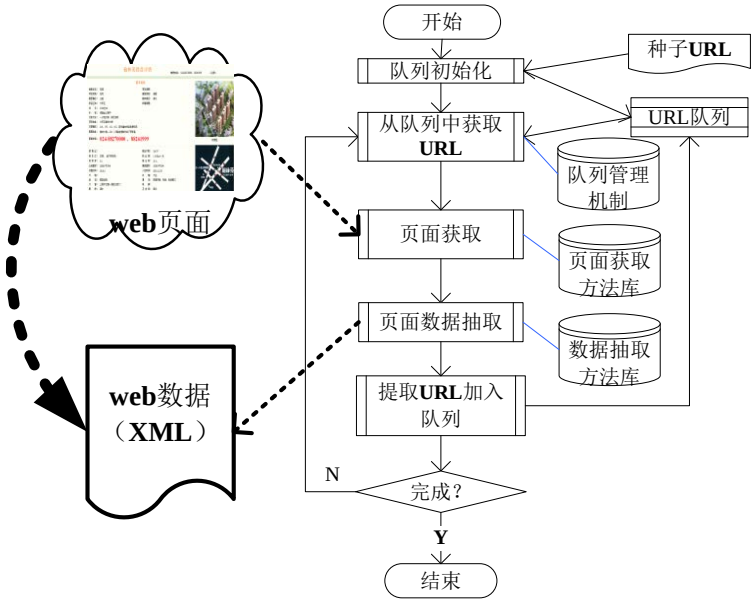

Figure 4. Process of web data extraction
图 4. Web 数据抽取流程图

1) 寻找领域种子 URL，
2) 依据 URL 队列获取网站的页面；
3) 获取页面中所需要的数据。

由于不同的网站采用不同的技术(包括网站设计开发技术、数据传输的方式和展现的形式等)，这决定了在实现 web 数据抽取的过程中，对于不同类型的网站必须采用的不同的数据抽取方法，这其中涉及到的关键技术包括：URL 队列管理机制、页面获取方法、页面数据的抽取方法等[7]-[10]。

URL 队列管理机制：由网站页面中的超链接关系建立网站页面的 DOM 树，依据网站页面中的超链接性质(静态链接和动态链接)采用不同页面搜索遍历方式(广度优先搜索遍历和深度优先搜索遍历)；

页面获取方法：包括单页面获取方法和多页面自动获取方法。在单页面获取方法中，必须依据网站页面采用的数据传输和展现的方式(比如 AJAX 技术)，获取所要抽取的数据值；在多页面自动获取方法中所要解决的问题包括，如何获取通过多个页面展现的一类 web 数据和通过页面只展现部分记录的后台数据库的全部记录。

数据的抽取方法：由于获得的网站页面数据包含大量的无用的噪声数据，数据的抽取方法就是解决如何依据页面的模式定位和解析目标数据，并以规范的格式存储等问题。

5.2. 领域增值服务

WEB 数据抽取与集成的目的是提供面向领域的增值服务，如图 5 所示，开发领域增值服务需要在建立领域集成数据库的基础上解决面向领域新要求的数据接口、外模式、服务模型和领域服务三个层面的问题。领域增值服务依据领域中不同客户群体的需求不同而不同。领域增值服务是建立在领域集成模式的外部模式基础上，通过服务模型来描述合实现的，同时为了实现上述服务对领域集成数据库的高效访问，需要建立统一的数据访问接口。

6. 系统实现-房地产大数据平台及综合应用系统

6.1. 系统概述

系统结合实际应用项目需求，依据不同城市房地产行业相关备案信息的网上发布情况，开发的基于房地产大数据平台的城市地产信息资源综合应用系统。

系统结合实际应用需求，依据不同城市房地产相关备案信息的网上发布和各城市发布网站的技术特点，将发布网站分为不同类型，并建立对应不同类型网站数据抽取方法库；建立统一的城市房地产数据集成模式和各城市发布网站对应的 web 数据模式；建立了不同网站的 web 数据模式到集成模式的模式映射规则库；建立不同城市网站抽取的 web 数据到集成数据库的集成方法库；结合领域实际需求，实现了面向政府监管部门的市场监测服务、面向开发商企业的决策支持服务、面向银行和税务部门的房产评估服务、面向公众的查询服务，以及面向开发商、经纪人和购房者的网上交易服务等。

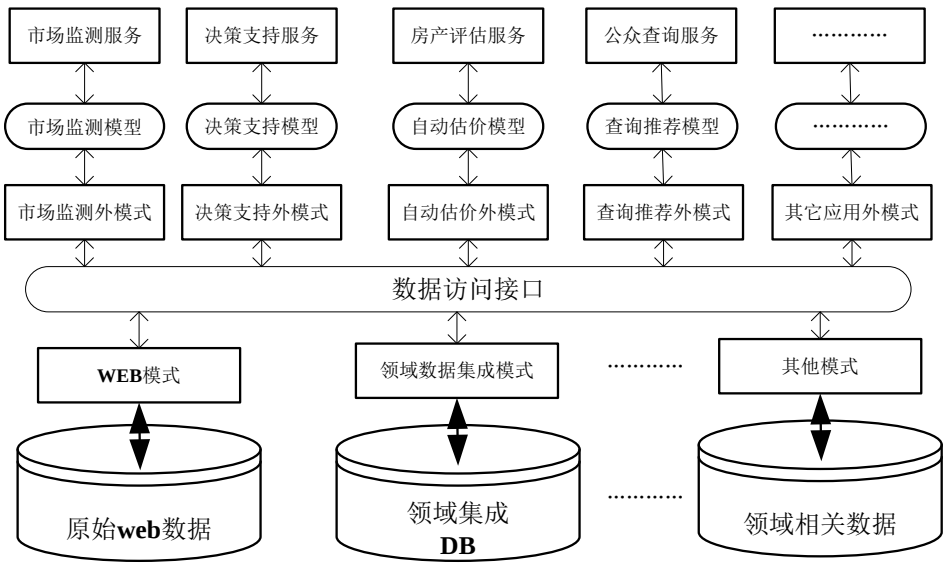

Figure 5. Hierarchy of domain value added service
图 5. 领域增值服务层次结构

6.2. 模式建立及映射

系统依据房地产领域的信息需求，结合各个城市商品房备案数据发布系统的信息特点，建立了城市房地产大数据平台的集成模式和各个城市商品房备案数据发布网站的 web 模式，并建立了城市发布网站的 web 模式到大数据平台的集成模式的映射关系。构成城市房地产大数据平台的逻辑结构。由于篇幅的关系，这里只讨论大数据平台中的有关商品房的核心数据，其中包括不同城市的楼盘、楼栋和户信息。

领域集成模式，包括：领域属性集(楼盘信息属性集、楼栋信息属性集和户信息属性集)、领域实体集(楼盘信息实体、楼栋信息实体和户信息实体)、领域约束集(一个楼盘包含多个楼栋，一个楼栋只能属于一个楼盘；一个楼栋包含多个户，一个户只能属于一个楼栋)；

网站 web 模式，对每一个城市商品房备案数据发布网站中的楼盘、楼栋和户信息的页面建立的对应的 web 模式，即，包括楼盘信息 web 模式、楼栋信息 web 模式和户信息 web 模式。

web 模式到集成模式映射，对每一个城市的楼盘信息 web 模式、楼栋信息 web 模式和户信息 web 模式，首先建立与领域集成模式中属性集的属性映射规则，然后建立到领域集成模式中楼盘信息实体、楼栋信息实体和户信息实体的实体映射规则。

6.3. 数据抽取与集成实现

6.3.1. 数据抽取实现

目前，系统已实现 40 余个城市商品房备案信息发布网站有关数据的定期抽取与更新(如表 1 所示)，这其中涉及的关键问题包括：网站类型分析和抽取方法库的建立；分布并行快速抽取的实现机制；防爬虫、图形数据识别、查询界面验证码等问题的解决。

6.3.2. 数据集成实现

数据集成所要解决的问题是将已经抽取完的每个城市的楼盘、楼栋和户数据装载到领域集成库中并保证数据的一致性。通过数据预处理和数据装载两阶段来实现。

数据预处理，依据属性映射规则，分别对每个城市的抽取的楼盘信息、楼栋信息和户信息的抽取数据中的每条记录的每个数据项进行预处理，这其中包括：数据项的类型转换、异常值处理和空值处理等操作。

数据装载，依据领域集成模式中的实体约束集合和模式映射中的实体映射规则，分别对每个城市已预处理完的的楼盘信息、楼栋信息和户信息的所有记录分别装载到集成数据库中楼盘信息表、楼栋信息表和户信息表中，并保证数据的一致性。算法流程如图 6 所示。

Table 1. Statistics data of cities have been extracted
表 1. 已实现的城市发布网站抽取数据的信息量统计

城市	许可信息	楼盘	楼栋	户型
北京	7610	2212	12,348	881,294
上海	17,404	5404	307,450	3,512,295
广州	8927	1321	63,231	1,669,550
深圳	2908	1561	10,430	1,140,545
沈阳	6221	1567	30,733	1,945,833
长春	3033	748	15,667	943,307
哈尔滨	1116	539	9257	885,683
。。。	。。。	。。。	。。。	。。。
总计(40 个城市)	160,651	35,926	772,732	31,019,981

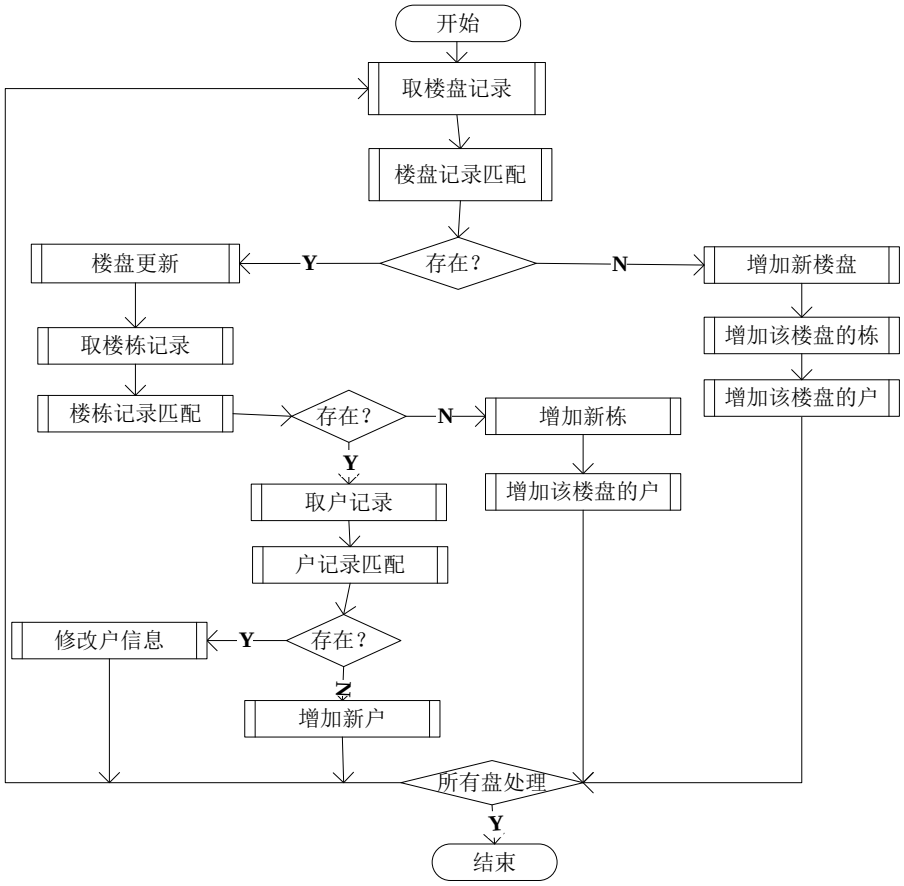

Figure 6. Process of page data extraction
图 6. 页面抽取数据装载流程图

7. 结束语

本文通过对现有的 Web 结构化数据抽取方法的分析，结合领域 Web 数据的分析结果和领域应用需求，从提供面向领域信息增值服务角度，提出一种面向领域的 Web 数据抽取与集成架构、Web 数据模型与 Web 数据模式、领域数据模型和领域数据模式；并给出 Web 数据模式与领域数据模式的映射方法、数据层次上的集成方法、web 数据抽取和领域增值服务的实现方法；结合房地产领域的实际需求，开发了房地产信息平台及综合应用系统。目前该研究领域需要进一步解决的问题是如何在突破网站的防爬虫技术、图形数据识别、查询界面验证码等问题的基础上，高效地实现面向领域的大规模非结构化 Web 数据的抽取与集成。

基金项目

辽宁省自然科学基金项目(2014020068)。

参考文献 (References)

[1] Cafarella, M.J., Halevy, A., Wang, D. Z., Wu, E. and Zhang, Y. (2008) WebTables: Exploring the Power of Tables on the Web. *Proceedings of VLDB*-08, **1**, 538-549. http://dx.doi.org/10.14778/1453856.1453916

[2] Liu, B. (2013) Web Data Mining [M]. 俞勇，薛贵荣，韩定一，译. 北京: 清华大学出版社, 2013.

[3] Volkovs, M., Chiang, F., Szlichta, J. and Miller, R.J. (2014) Continuous Data Cleaning. CDE, 244-255.

[4] Geerts, F., Mecca, G., Papotti, P. and Santoro, D. (2014) Mapping and Cleaning. ICDE, 232-2243.

[5] 李贵, 张淼. 基于领域模型的 Web 数据抽取与集成[J]. 微电子学与计算机, 29(9): 152-156.

[6] 马安香, 张斌, 高克宁, 齐鹏, 张引. 基于结果模式的 Deep Web 数据抽取[J]. 计算机研究, 46(2): 280-288.

[7] Gatterbauer, W., Bohunsky, P., Herzog, M., Krüpl, B. and Pollak, B. (2007) Towards Domain-Independent Information Extraction from Web Tables. *Proceedings of WWW*-07, Banff, 8-12 May 2007, 71-80. http://dx.doi.org/10.1145/1242572.1242583

[8] Sheng, C., Zhang, N., Tao, Y.F. and Jin, X. (2012) Optimal Algorithms for Crawling a Hidden Database in the Web. *Proceedings of VLDB*, **5**, 1112-1123. http://dx.doi.org/10.14778/2350229.2350232

[9] 田建伟, 李石君. 基于层次树模型的 Deep Web 数据提取方法[J]. 计算机研究与发展, 2011, 48(1): 94-102.

[10] 寇月, 李冬, 申德荣, 于戈, 聂铁铮. D-EEM-一种基于 DOM 树的 Deep Web 实体抽取机制[J]. 计算机研究与发展, 2010, 47(5): 858-865.

[11] Wang, R. and Cohen, W. (2008) Iterative Set Expansion of Named Entity Using the Web. ICDM.

[12] Pantel, P., Crestan, E., Borkovsky, A., *et al.* (2009) Web-Scale Distributional Similarity and Entity Set Expansion. *Proceedings of EMNLP* 2009, Singapore, 6-7 August 2009, 938-947.

[13] 李贵, 陈韶刚, 等. 基于 Web 的实例扩展与属性值扩充方法[J] 计算机科学, 2014, 41(11A): 411-418.

[14] Dalvi, N., Rastogi, V. and Dasgupta, A. (2013) Optimal Hashing Schemes for Entity Matching. *Proceedings of the 22nd International Conference on World Wide Web*, Rio de Janeiro, 13-17 May 2013, 295-305. http://dx.doi.org/10.1145/2488388.2488415