文章编号:1001-5132（2008）02-0201-05

面向电子商务的Deep Web数据集成研究

张大吉

（宁波大学 图书馆，浙江 宁波 315211）

摘要：随着 Internet 信息的迅速增长与网络技术的不断进步，越来越多的数据库可以通过网络查询接口直接访问，包含这种类型数据库的 Web 被称为 Deep Web. 互联网中的 Deep Web 数据存储量非常庞大，而其中大部分是电子商务数据. 本文提出了电子商务 Deep Web 数据集成系统架构，并对其中关键问题进行了介绍，包括 Deep Web 的发现、接口的抽取与集成、结果的抽取与整合等.

关键词：Deep Web；数据集成；电子商务；Web 数据库

中图分类号：TP393 文献标识码：A

近来，随着互联网络的迅猛发展，Web中所隐藏的巨量信息越来越受到人们的关注. 根据数据的分布状况，Web可以分为：Surface Web和Deep Web. Surface Web是指可以通过超链接或者传统网页搜索引擎访问到的网页、文件等资源. 它一般以静态网页构成为主. 而Deep Web可以简单的概述为那些难以通过普通搜索引擎发现的资源的集合. 主要包括存储在Web数据库里大量资源，需要通过动态网页技术才能访问. 据统计[1]，Deep Web的数据存储量是Surface Web的 400～550 倍. 2004 年互联网上的Web数据库已经达到了 450 000 多个，信息量超过了 200 000 TB，并且这个数据还在不断地增大. 另一方面，Deep Web后台的数据库一般为结构化的关系数据库，质量都比较高，因此通过Deep Web的数据集成来更有效地利用Deep Web丰富的数据是十分有意义的. 本文将从电子商务的角度研究Deep Web数据库集成的应用.

1 电子商务 Deep Web 数据集成系统架构

互联网中，电子商务的数据信息是巨大的，它们大部分储存在电子商务网站(淘宝、EBAY 等)的 Web 数据库中，用户只能通过某个网站的查询接口(如图 1)查询并获得其 Web 数据库的信息. 本文给出了一个电子商务 Deep Web 数据集成系统架构，通过该系统用户可以查询到网络中的电子商务 Deep Web 数据(商品信息等)，从而选择最合适自己的商品进行交易.

在这个电子商务 Deep Web 数据集成系统中，主要有以下 3 个关键的步骤：(1)发现电子商务 Deep Web；(2)查询接口抽取与集成；(3)结果抽取与整合.

收稿日期：2007-12-10. 宁波大学学报（理工版）网址：http://3xb.nbu.edu.cn
基金项目：国家自然科学基金（0773064）；浙江省自然科学基金（Y105625）.
作者简介：张大吉（1981－），男，四川彭州人，助理馆员，主要研究方向：数据挖掘及数据库. E-mail: beyond@nbu.edu.cn

图 1　淘宝网和拍拍网的查询接口

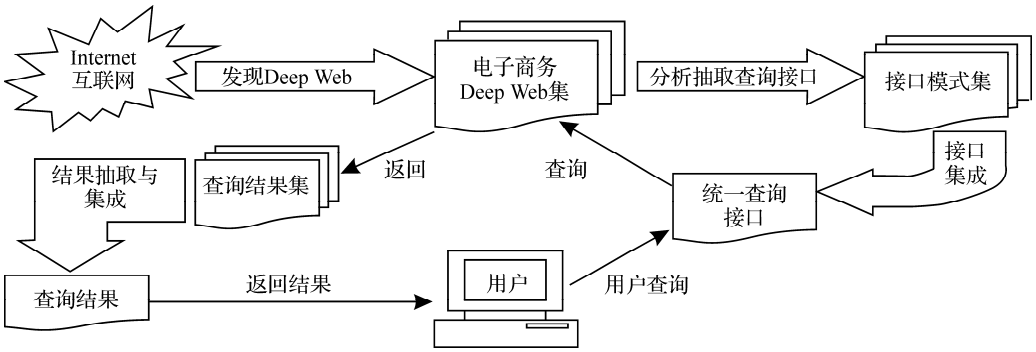

图 2　整体架构

图 2 给出了该系统的整体架构.

2　电子商务 Deep Web 的发现

在这个模块里需要完成 2 个功能，首先需要发现含有电子商务信息的 Deep Web 网站，然后再从发现的网站中获取 Deep Web 的查询接口.

通过对网络中某些网站的人工辨别，可以很准确地发现Deep Web网站地址，并获取其查询接口，这对于小规模的数据集成来说是一个很有效的解决方案. 但是对于电子商务来说，信息的及时性和规模性是其重要基础，手工获取Deep Web网站不能很好地解决这个问题，因此采用自动发现Deep Web网站的技术. 自动发现电子商务Deep Web有如下步骤：(1)获取某个网站地址；(2)判断其是否为电子商务Deep Web. 基于以上 2 个步骤，我们一般有 2 种解决办法，第 1 种是遍历所有互联网中的IP地址[3]，通过这种方法可以获取所有存在Http服务的站点地址，然后再进行电子商务Deep Web的判断. 但是，遍历网络总几十亿个IP的工作量是十分巨大的，代价也过于昂贵，因此一般不采用这种方法. 第 2 种是首先建立一个电子商务本体知识库，这个知识库可以根据网络上用户感兴趣的领域人工建立，也可以通过应用程序输出；然后在这个知识库的基础上结合搜索引擎(Google，Baidu)来获取网站地址，通过这个方法获取的网站地址质量普遍较高.

Deep Web后台数据库的唯一入口是查询接口，因此发现这些接口是判断的前提，确定为Deep Web之后，这些接口又是数据集成的重要组成部分. 根据研究表明：94%的查询接口深度值为 3，因此只需要遍历网站的浅层页面即可获得其查询接口[4]. 如何将查询接口表单从一般表单中分离出来并判断其是否为电子商务Deep Web呢？我们可以通过上文提到的电子商务本体知识库和普通查询接口的基本特征[5]来描述产生电子商务Deep Web查询接口特征，然后在这些特征的基础上利用文献[6]中的算法建立一棵决策树，通过这棵决策树找出真正的电子商务Deep Web查询接口. 该模块的整体框架如图 3 所示.

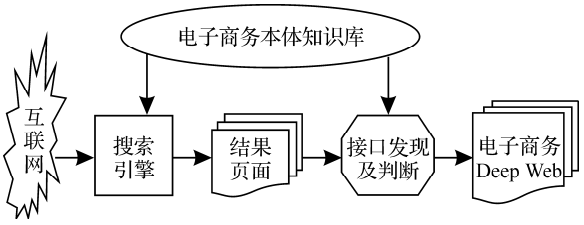

图 3　发现 Deep Web 途径

3　查询接口的抽取与集成

用户通过查询接口来获得 Deep Web 的数据信息，一个查询接口物理上通常包括一些 HTML 元素，比如：Textbox, Button, Checkbox, Radio 等等，还包括一些有语义的文本标签. 而在逻辑上，一个查询接口包括了很多的逻辑属性值，如图 1 中的“淘宝网”查询接口就包括了“关键”、“类别”等属性；而每个不同的逻辑属性又包括了一些特定的元数据信息，比如：属性名称、数据类型、数据范围、数据单位、默认值、查询条件等，如图 1 中“拍拍网”的“价格范围”属性就包括了这些元数据信息：名称——价格，数据类型——货币，单位——元，查询条件——最大与最小范围查询等等. 我们把接口的逻辑形式称为接口模式.

从不同的接口模式中抽取出特定的逻辑属性，并根据相关性把他们集合成为一个通用的逻辑属性，这是查询接口抽取的主要目的. 而通过对通用逻辑属性的集成，将得到一个统一的查询接口.

3.1　查询接口的抽取

一个接口模式的逻辑属性可以用一组特定的元数据信息来确定，而这些信息都是分散在接口的物理元素中的，因此有必要通过搜寻这些元数据信息，对某个逻辑属性进行确切的定义，从而准确地表达这个接口模式.

在抽取查询接口的过程中，引入了接口表达式(Interface Expression，简称IEXP)[2]的概念，通过对接口表达式的改进，可以很方便地完成对接口模式的抽取. 如图 1 中淘宝网查询接口的接口表达式为：

I {TE|T|TEEEE|TE|TE|TE|BB}

其中：“T”(Text-label)代表接口中的文字标签，“E”(Element)代表接口中的 HTML 元素，但不包括按钮，按钮用“B”(Button)来表示，“|”表示表格的一行或者换行符号.

接口表达式可以对网络中大部分查询接口做出形象的描述，它对我们下一步接口抽取起着重要的作用. 通过对接口表达式的分析抽取，可以得到该接口模式的逻辑属性，下面给出了 2 种抽取方法：

(1) 基于 T(文本标签)的抽取方法(Text-label based，简称 TB). 对于一个接口模式，找到其中的所有文本标签(IEXP 中的 T)并逐个进行如下分析：在某个标签 T1 的同一行或者下面临近行中往下找寻与之相邻的 HTML 元素(E)，直到另一个文本标签 T2 为止；把这些 E 与 T1 合并成一组进行启发式的分析(比如比较文本标签与 HTML 元素的名称)，将不匹配的 E 抛弃；如果最终没有任何 E 与 T1 匹配，那么 T1 将被抛弃，否则 T1 和与其相匹配的{E11, E12, E13…}构成一个逻辑属性 A1.

(2) 基于 E(HTML 元素)的抽取方法(Element based，简称 EB). 该方法与上述方法刚好相反，首先找到 E1，然后往上找寻 T，进行启发式分析，如果匹配则停止找寻并将他们合并逻辑属性 A1，否则继续往上找寻 T，直到另一个 E2.

我们手工获取了网络中的 100 个查询接口，然后用 2 种方法分别对这些接口进行分析抽取测试，实验数据为：在总抽取的 100 个查询接口中，TB 抽取成功 74 个，EB 抽取成功 81 个，均抽取失败 7 个.

通过以上方法对接口的逻辑属性进行抽取后，通过进一步分析，可以把一个逻辑属性表示为：A[Name, Type, Range, Layout…]，其元信息包括属性名称、数据类型、值域、排列位置等等. 于是，一个查询接口经过抽取和分析后可以形象地表示为：I{A1, A2, A3, …, A*n*}，Ai 代表接口的若干逻

辑属性.

3.2 查询接口的集成

对于查询接口的集成有 2 个步骤：首先将不同接口中的特定逻辑属性集成为通用的逻辑属性；然后将这些通用的逻辑属性集成为一个统一的接口.

在不同的查询接口中，语义相似或相同的属性可能会被表示成不同的模式，比如不同的文字标签，不同的 HTML 元素格式，不同的排列布局等等. 为了得到通用属性，本文提出了 2 种集成的思路：(1)基于本体知识相关库的集成，通过建立知识相关库，对文本标签、属性名称进行相关性判断，然后集成. 这种方法的结果比较准确，但成功率不高，容易造成资源浪费. (2)基于属性模式的集成，通过对不同接口的逻辑属性 Ai [Name, Type, Range, Layout…]中的元信息(名称、类型、值域、排列位置等)进行语义关系分析、模式匹配，然后确定出不同属性间的相似度，最后根据相似度进行集成. 这种方法的效率和成功率都较高，准确率也比较乐观.

在通用属性的集成过程中，需要增一个重要的元信息：集成度. 通用属性的集成度是对所有被集成的逻辑属性的量化反映，另外我们还需要建立每个逻辑属性到该通用属性的映射信息(域名、名称等). 流程如下：

```
while (CanBeIntegrated(At,Ai))  //判断逻辑属性 Ai 是否能集成入通用属性 At
{
    At=Integrator(At,Ai);       //集成 Ai 到 At
    At.integration++;           //At 的集成度加 1
    At.mapping=Mapping(At,Ai); //建立 At 到 Ai 的映射信息
    i++;          //循环判断下一个逻辑属性
}
```

通过以上对通用属性的集成，可以很方便地得到统一接口. 将集成度 $> n$ 的通用属性挑选出来作为该统一接口的逻辑属性，并根据它的映射信息得到该属性与其他接口的关系. 最后根据各个通用属性的排列位置元信息对该属性进行位置排列.

4 结果的抽取与整合

用户在统一的查询接口提交查询，电子商务 Deep Web 返回的结果主要是通过 HTML 语法的页面来展示，而每个 Deep Web 返回的页面结构都是不同的，因此需要通过结果的抽取获得有价值的内容，并对这些内容进行整合，然后返回给用户. 在这个模块中，关键问题是如何从查询返回页面大量数据中抽取有价值的结果. 目前在这个研究领域国内外学者已经开展了大量的研究工作，利用文献[7]中介绍的方法，通过建立 DOM 树可以自动地完成查询结果的抽取工作.

该方法提出的 MDR 算法，能够比较准确地完成多记录页面的抽取. 它的主要思想是通过分析页面的 HTML 结构建立 HTML 标签树，然后确定通过比较标签树中的节点路径或者其结构信息，发现代表查询结果的节点并进行标记. 最后对树中这些节点数据进行抽取. 我们将抽取的数据用 XML 形式保存，通过对所有 XML 数据的整合，将查询结果以统一的格式呈现给用户.

5 结束语

本文以电子商务为模型提出了 Deep Web 数据集成的应用解决方案，并对其中的关键问题进行了讨论，包括 Deep Web 的发现，接口的抽取与集成，结果的抽取与集成等. 随着互联网技术的不断发展，网络中 Web 数据库的数目不断增加. 近几年来，国内外专家在这个领域中做了大量的工作，但是关于 Deep Web 的研究仍处于起步阶段，后续还有许多问题需要我们更深入的研究，比如 Deep Web 多数据源的聚类与分类等等.

参考文献：

[1] Bergman M. The deep web: Surfacing hidden value[J]. Journal of Electronic Publishing, 2001, 7(1):1-17.

[2] He Hai, Meng Weiyi, Yu Clement, et al. Constructing interface schemas for search interfaces of web databases[C]//6th International Conference on Web Information Systems Engineering (WISE05), New York. 2005:29-42.

[3] Kevin Chenchuan, He Bin, Li Chengkai. Structured databases on the web: Observations and implications[J]. SIGMOD Record, 2004, 33(3):61-70.

[4] He Bin, Patel Mitesh, Zhang Zhen. Accessing the deep web: A survey[J]. Communications of the ACM, 2007, 50 (5):94-101.

[5] Cope J, Craswell N, Hawking D. Automated discovery of search interfaces on the web[C]//Proceedings of the 14th Australasian Database Conference (ADC 2003), Adelaide, 2003.

[6] J Quinlan. C4.5: programs for machine jearning[M]. San Mateo: Morgan Kaufmann Publisher, 1993.

[7] Zhai Y, Liu B. Web data extraction based on partial tree alignment[C]//Proceedings of the 14th International World Wide Web Conference Committee (IW3C2), China, 2005.

[8] Marcus P. Knowledge discovery resources 2007[EB/OL]. [2007-2-12]. http://www.llrx.com/features/knowledgedis-covery.htm.

[9] Marcus P. Deep web research 2007[EB/OL]. [2006-12-17]. http://www.llrx.com/features/deepweb2007. htm.

The Research of Deep Web Data Integration for E-commerce

ZHANG Da-ji

(Library of Ningbo University, Ningbo University, Ningbo 315211, China)

Abstract: As the amount of information on the web increases rapidly, more and more databases are becoming Web accessible through form-based search interfaces, and the web with this kind of database is called “deep web”. The size of the deep webs in internet is immense, many of which are E-commerce sites. In this paper, the architecture of the deep web data integration system for E-commerce is presented, and the key components in the system are introduced, among which are Deep Web discovery, query interface extraction and integration, web data extraction and merging.

Key words: Deep Web; web data integration; e-commerce; web database

CLC number: TP393　　**Document code:** A

（责任编辑　史小丽）