一种深度网络数据库集成技术研究

杨宏伟　马永征　钱芳
中国科学院计算机网络信息中心，北京 100190

摘　要： 近来研究表明，Deep Web提供的高质量专业数据信息对e-Science环境是极为有价值的。本文就如何集成深度网数据库技术进行了研究与探讨，包括建立有效的Deep Web爬虫、匹配Schema自动填写查询表单、集成深度网数据库查询接口以及建立统一用户查询界面等相关技术。

关键词： 深度网；深度网数据库；深度网爬虫

An Integration Technique of the Deep Web Database

Yang Hongwei，Ma Yongzheng，Qian Fang
Computer Network Information Center Chinese Academy of Sciences，Beijing 100190,China

Abstract: According to recent studies,the content provided by many hidden web sites is often of very high quality and can be extremely valuable to many users who work under e-Science environment.This paper reviewed the technology for integration Deep Web Database, including building an effective hidden web crawler, matching form schemata and automatically filling out forms.With those obtained query interfaces, they can be integrated to obtain a unified interface which is given to users to query.

Keywords: Deep web;Hidden web database;Deep web crawler

＊ 资助项目：863计划科学数据网格及科研应用系统（课题号2006AA01A120）

1. 引 言

Internet已经成为当今社会人们获取信息的主要来源，尤其是数据库技术与网络技术的结合，使Internet拥有了最为巨大的信息量，进而衍生出深度网（Deep Web）。最初由Dr. Jill Ellsworth于1994年提出，指那些由普通搜索引擎难以发现其信息内容的web页面。2001年，Christ Sherman等定义Deep Web为：虽然通过互联网可以获取，但普通搜索引擎由于受技术限制而不能或不作索引的那些文本页、文件或其它通常是高质量、权威的信息。据最近对Deep Web的调查[1]得到了以下有意义的发现：

当前Deep Web的规模为307,000个站点，450,000个数据库和1,258,000个查询接口，在2000-2004年间增长了3-7倍；

Deep Web广泛分布于几乎所有的学科领域；

Deep Web对于主流搜索引擎来说并不是完全不可见，大约有1/3的数据已经被覆盖；

Deep Web中的数据大多是结构化的；

尽管一些Deep Web的目录服务已经开始索引Web上的数据库，但是他们的覆盖率很小约为0.2%到15.6%；

Web数据库往往位于站点的较浅层，94%的Web数据库位于站点前3层。

由此可见，获取Deep Web数据是解决e-Science环境的用户获取所需要的数据信息重要问题之一。与此同时，深度网上的数据存在着即使是同一领域的数据在不同的数据源上用户检索接口所采用的技术都各不相同，用户为了检索某类数据不得不熟悉多个数据源的检索语法并要进行多次检索，检索的结果必须要用户自身进行分析以确定是否有价值。深度网的数据源集成问题目前已经成为构建e-Science环境的一个重要的环节，本文为解决深度网数据集成问题进行了一些研究与探索，提出了一种理想的集成模型并对其实现技术进行了初步阐述。

2. Deep Web数据集成框架

为了提高数据网数据利用的有效性，理想的方式是为用户提供一个数据集成系统，按照所属领域把数据源进行分类，每类数据源提供一个统一的访问接口，从而使用户能够按照所要检索信息的类别来选择相应的查询接口，由集成系统根据用户输入的内容向所集成的Deep Web数据资源发送检索请求，并对返回的检索结果进行排序，最终呈现给用户。

理想情况下用户通过集成检索界面输入查询信息；该查询由查询处理模块进一步处理，根据查询内容选择Deep Web数据源（可能是多个），而后根据不同的Deep Web数据源的要求转换用户查询内容以适应Deep Web数据源，向Deep Web数据源提交查询；检索的结果由结果处理模块进一步处理，由于返回的结果一般是HTML页面的形式需要提取数据，提取后的数据需要根据领域进行标注以解释其意义，合并所有数据结果呈现给用户。但是，建立深度网数据集成系统面临诸多挑战[2]：

数据源发现：由于深度网数据一般存储在后台数据库中，用户需要通过填写数据源提供的HTML表单内容后，由数据源产生查询命令，最终以动态生成HTML文件显示在浏览器上的方式返回检索结果。如何有效地从Internet上搜索标记这些能够检索深度网数据的HTML表单页面是个很难解决的问题。

为了建立统一的检索界面需要解决深度网数据源检索模式提取、匹配、合并这些棘手问题。对于同一语义的检索项，不同的深度网数据源采用不同的表示方式等等，这样的问题都需要通过各个不同深度网数据源的检索模式高度抽象，最终以统一的检索模式反映出所有数据源的特征才能达到建立统一检索界面的目的。

查询优化：深度网数据资

源以分布式的方式存在Internet上，完全是一种超大规模的异构体，用户的一次检索请求需要大量的相关深度网数据源进行合作完成，优化查询决定了用户查询请求是否能够完成。

结果合并：用户的一次检索请求会有成百上千个深度网数据源响应并返回结果，去除重复、结果排序等问题的解决决定了系统的可用性。

系统维护：由于深度网数据源是动态变化的，集成系统能否及时反映这些变化，如数据项的增减等，也必将决定系统的灵活性、适应性。

2.1 集成模型

为了更好地解决深度网数据集成问题，根据其特点本文设计了集成模型如图1，深度网搜索引擎按照专业领域来构建，不同领域呈现不同的搜索界面，用户根据检索的领域知识来选择检索界面，搜索引擎根据用户检索请求，通过全局模式（Global Schema）映射选择与用户检索请求相关的深度网数据检索接口，形成检索入口页簇（Entry Pages），通过检索接口转换模块（Query Interface）向相关的深度网数据源发出检索请求；所有检索得到的结果页面最终返回到搜索引擎，搜索引擎根据全局模式对检索结果进行标注、归并、排序并呈现给用户。

2.2 基于本体的聚焦爬虫

考虑到Deep Web的唯一入口点是搜索表单，而传统的搜索引擎搜索的主要对象是HTML页面，因而利用传统搜索引擎的爬虫技术是一种很自然的选择。Focused Crawler概念最初提出的目的就是在互联网中抓取特定主题的网页。文献[3]是具有代表性的聚焦爬虫的早期研究之一，目前大多数的聚焦抓取都采用了类似的工作流程。

基于本体的聚焦爬虫根据特定的抓取目标。有选择的访问相应网页和链接，获取所需要的语义信息。我们将表单表示为二元组QI(element，domain)形式，并尝试通过语义标注、页面布局等信息确定表单的输入数据模式[4]。领域本体由不同的概念、实体及其之间的关系以及与之对应的叙词表组成。网页中的关键词在通过与领域本体对应的词典作规范化转换之后，进行计数和加权处理，算出与所选领域的相关度。对规范化后的词进行加权时，根据本体的概念层次，离核心概念越近的权重越高。然后分析数据源接口集SII中页面包含的比较丰富的语义信息，通过这些语义信息来帮助判断一个Form表单是否为查询接口QI，这样可以达到比传统的基于关键字的分类分析算法具有更高的准确性和效率。另外，即使初始URL与领域不

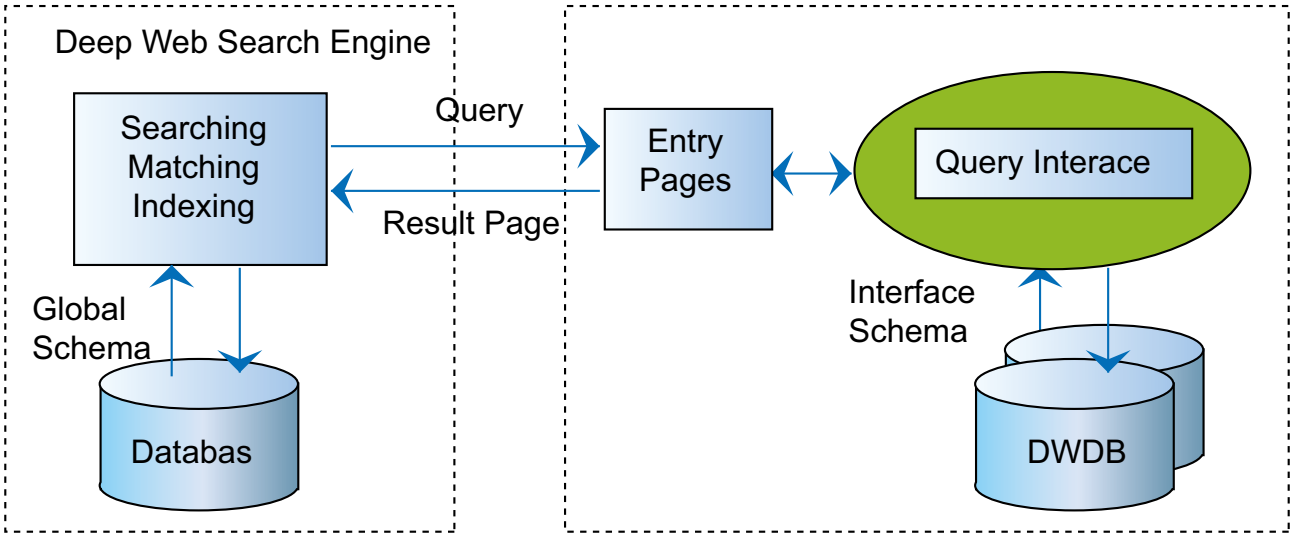

图1 Deep Web数据集成模型

直接相关，也具有较好的抗干扰能力，并逐渐趋近于主题相关的抓取路径。为此采用了本体词典和相关领域叙词表来表示信息，领域信息定制使爬虫可以有目的地从网页上抓取Deep Web数据，返回结果具有一定结构的数据信息及附带的若干网页URL等。通过词典进行字词扩展，领域信息和元搜索技术结合，使得聚焦爬虫抓取过程中，可兼顾准确率和覆盖率。

2.3 查询接口模式的抽取

对于利用爬虫所获取的含有查询接口的页面还要进一步进行查询接口模式抽取。查询接口的模式是一组领域相关的属性集合，通过对其中若干属性的赋值形成一个对该查询接口所代表Web数据库的查询。对查询接口模式的抽取可以获得一个查询接口的查询能力，查询接口的模式可以被看作是建立在对应Web数据库上的一个视图。对查询接口模式的抽取是指对查询接口属性的获取与分析。对查询接口模式的抽取主要目的是为了下一步的Web数据库分类和查询接口集成，其关键是把查询接口所包含的各个属性准确地抽取出来。文献[5]以文法分析的方式来完成对查询接口模式的抽取。

2.4 查询接口分类策略

数据源聚类算法基本思想是基于查询接口特征相似度来聚类相似的接口表单。两个查询接口的特征相似度被定义为它们表单术语加权相似度和功能术语加权相似度之和。

定义1[6]两个Deep Web数据源查询接口dwi_1和dwi_2的相似度值为：

$similarity(dwi_1, dwi_2)=W_1+FormSim(dwi_1, dwi_2)+W_2*RegularSim(dwi_1, dwi_2)$

其中$FormSim(dwi_1, dwi_2)$为两个查询接口的表单术语相似度，$RegularSim(dwi_1, dwi_2)$为两个查询接口的功能术语相似度，功能术语可以从查询接口对象中数据项对应部分提取出来。两个向量之间的相似度可以使用Cosine相似度函数来计算，W_1和W_2分别为两种术语对应的权重系数，可以通过实验获得。

基于Deep Web查询接口的分类策略是先抽取查询接口页面特征，接着根据查询接口特征对查询接口进行聚类、自动为聚类委派概念名。然后使用已有的全局分类模式结构来对聚类进行合并和分类。查询接口聚类过程中的聚类是独立于应用领域的，而一个特定的应用领域有时候需要合并某些基本的聚类。将一个聚类委派到合适的概念层次上，首先要为每个聚类自动产生描述性术语，同时给已知主题层次上的概念自动产生描述性的术语，然后通过比较这两个术语的相似度来确定某聚类的概念。

2.5 查询处理

查询接口的集成是为了给用户提供一个对属于同一个领域的Web数据库统一的访问途径，而对Web数据库的访问方式主要是通过查询接口，因此对Web数据库集成重要的一步就是查询接口的集成。集成的查询接口合并了同一领域的查询接口集合中表示同一语义的属性，保留了一些查询接口中特定的属性，并尽可能地保持该领域查询接口的结构特征和属性的顺序性。如果把各个被集成的查询接口看作Web数据库的一个本地视图的话，那么集成的查询接口就是建立在这些本地视图之上的全局视图。

查询结果处理模块将各个Web数据库返回的结果抽取并合并到一个统一的结构化的模式下，该部分包括结果的抽取、结果的注释和结果的合并。查询结果的抽取是指从Web数据库返回的结果页面中抽取出真正的查询结果。结果的注释是指由于抽取的结果通常缺少语义，因此要为缺少语义的数据项进行语义注释。查询结果的合并是指把从各个Web数据库得到的查询结果进行有效的合并去重，存储在一个统一的模式下。

3.系统实现的软件体系结构

根据前面设计的集成模型，我们把系统最终实现过程分为四个阶段：聚焦爬虫获取Form页面阶段，提取深度网数据源检索接口并形成全局模式库阶段，构建 ≫

面向领域集成检索接口阶段与获取查询结果归并排序阶段。

（1）系统中的Form爬虫根据领域专业词库的指导，从初始URL集读取起始点网站的URL，Form爬虫对起始站点的Web页面进行解析，若发现Web页面中包含有表单，则利用Form解析器将表单转换成DOM对象树。

（2）对上述的DOM对象树进行识别，深度网查询接口根据专业领域词库向获取的表单进行填写查询数据，若输出为专业领域相关的结果，则判定该DOM对象树有效，调用系统中的接口解析器对DOM对象树进一步解析，把查询接口的各个组件、隐藏参数、请求路径以及相关元数据等转换为系统中的全局模式的一部分，否则销毁该DOM对象树。

（3）Form爬虫在对一个Web页面的表单提取完之后，抽取页面中的超链接加入到未访问队列中，把当前URL加入到已访问队列中。重复上述过程。

（4）随系统中全局模式库的不断扩大，系统采用机器学习的方式生成面向专业领域的用户统一查询接口界面。

（5）系统中的全局模型库不断调用查询生成器构造查询命令，根据检索结果，模式库不断调整检索接口的数据项，不断有效统一查询接口。

4. 结束语

随着Web数据库在Web中不断大量的涌现，对Web数据库进行大规模集成的研究成为一个非常迫切的问题。至今人们在Deep Web领域已经作了大量的研究，所提出的Deep Web数据集成系统有文献[7, 8]，但它们只是属于研究性的原型系统，因此确切地说至今还没有一个真正可以作为实际应用的Deep Web数据集成系统。构建e-Science离不开Internet，更离不开Deep Web，因此如何把现有的Deep Web领域的工作成果应用到e-Science环境下，以及发展更有效的发掘Deep Web数据资源的新方法、新工具既具有现实意义也具有理论研究价值。

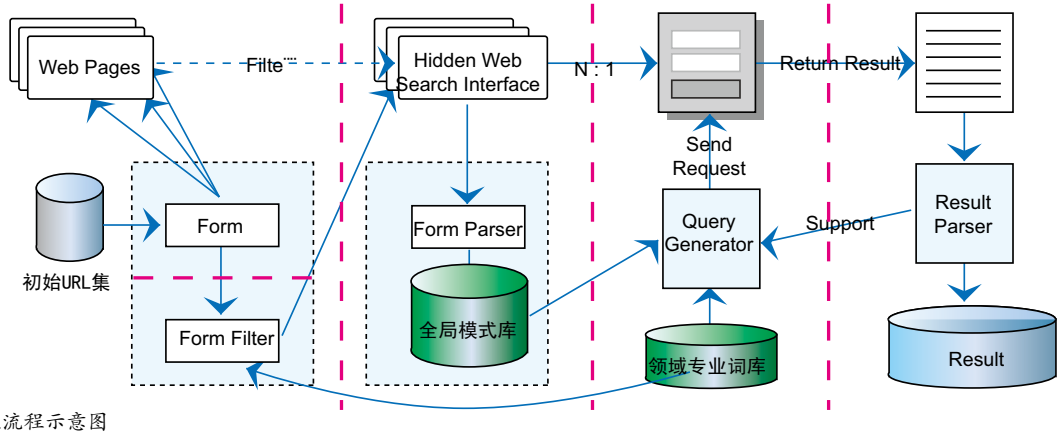

图2 实现处理流程示意图

参考文献：

[1] K.C.Chang,B.He,C.Li,M.Patel,Z.Zhang.Structured Databases on the Web: Observations and Implications. SIGMOD Record,2004,33(3)：61-70.

[2]刘伟,孟小峰,孟卫一.Deep Web数据集成研究综述,计算机学报,2007(9)：1476-1489.

[3] Chakrabarti S,van den Berg M,Dom B. Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. In: The 8th Intl. World Wide Web Conference,Toronto,Canada,1990.

[4] Lage J P,da Silva A S,Golgher P B,et a1.Automatic generation of agents for collecting hidden Web pages for data extraction[J]. Data&Knowledge Engineering, 2004,49: 177-196.

[5] Zhang Z. He B, Chang K C. Understanding Web query interfaces：Best-effort parsing with hidden syntax// Proceedings of the 23rd ACM SIGM OD International

Conference on Management of Data.Paris,2004:107-118.

[6] 王兵,王轲.Deep Web数据源聚类与分类.计算机与现代化,2007(8):36-40.

[7] He, H., Meng, W., Yu, C., Wu, Z. WISE-integrator: An automatic integrator of Web search interfaces for e-commerce// Proceedings of the 29th International Conference on Very Large Data Bases. Berlin, 2003: 357-368.

[8] Chang K C, He B, Zhang z. Toward large scale integration: Building a MetaQuerier over databases on the Web//Proceedings of the 2nd Biennial Conference on Innovative Data Systems Research. Asilomar, 2005: 44-55.

[9] http://www.green500.org/.

[10] Satoshi Matsuoka. The Road to TSUBAME and Beyond. High Performance Computing on Vector Systems 2007. Doi: 10.1007/978-3-540-74384-2-19.

[11] Minoru Nomura. Petascale Computing Trends in Europe. Quarterly review, 2008, 27(4).

[12] http://www.irp.oist.jp/hpc-workshop/presentations/Tadashi_Watanabe.pdf.

[13] http://nsf.gov/pubs/2008/nsf08592/nsf08592.htm.

[14] Rick Stevens. The LLNL/ANL/IBM Collaboration to Develop BG/P and BG/Q. http://www.sc.doe.gov/ascr/ASCAC/Stevens-ASCAC-March20061.pdf.

[15] http://www.nsf.gov/news/news_summ.jsp?cntn_id=106791.

[16] Streitz, F., et al. 100+ TFlop solidification simulations on BlueGene/L. In Proceedings of IEEE/ACM Supercomputing '05. http://sc05.supercomputing.org/schedule/pdf/pap307.pdf.

[17] Francois Gygi, et al. Large-Scale Electronic Structure Calculations of High-Z Metals on the BlueGene/L Platform. In Proceedings of IEEE/ACM Supercomputing '06. http://sc06.supercomputing.org/schedule/pdf/gb104.pdf.

[18] J.N. Glosli, et al. Extending Stability Beyond CPU-Millennium: Micron-Scale Atomistic Simulation of Kelvin-Helmholtz Instability. In Proceedings of IEEE/ACM Supercomputing '07. http://sc07.supercomputing.org/schedule/pdf/gb109.pdf.

[19] http://lca.ucsd.edu/software/enzo/nightly/userguide/.

收稿时间:2008年12月20日

作者信息

杨宏伟

中国科学院计算机网络信息中心，助理研究员。主要研究方向为下一代因特网和网络安全。

马永征

中国科学院计算机网络信息中心，博士，副研究员。主要研究方向为海量数据处理及分析和协同科研环境关键技术。

钱芳

中国科学院计算机网络信息中心，助理研究员。主要研究方向为网格技术、软件自动测试。