基于Deep Web的信息采集系统

王冉冉,王 刚,黄青松
(昆明理工大学 信息工程与自动化学院,云南 昆明 650051)

摘 要:随着互联网技术的迅速发展,大量结构化的高质量信息被埋入网络,却无法被传统的搜索引擎检索到,进而难以被挖掘利用。针对这一现象,提出了基于Deep Web的信息采集系统,设计了基于Web的查询方式,并结合数据挖掘的相关技术,获取并挖掘深网信息资源,解决传统手工采集信息的弊端,提高系统的使用效率,避免人工搜集时间和费用上的开销,降低成本,便于维护。并且正在云南省大型仪器协作共用网络平台的建设中尝试实现这个子系统的设计。

关键词:Deep Web;信息采集;查询接口;数据挖掘
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2007)10-0171-03

An Information Extraction System Based on Deep Web

WANG Ran-ran, WANG Gang, HUANG Qing-song
(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650051, China)

Abstract: With the rapid development of Internet technology, a large amount of structured and high-quality information is embedded into Internet. However, the information cannot be retrieved by traditional search engine and then it is difficult to mine and make full use of it. In view of this phenomenon, presents a system based on the deep Web information extraction, designs a query schema based on the Web, and combines some relevant technology of data mining. As a result, can get and mine the information which is in the Deep Web. At the same time, it resolves the traditional drawback of collecting information artificially, enhances the efficiency of the system, avoids the expenses on collection time and the expense, reduces the cost and maintains easily. And it has been designing in the Yunnan province scientific instrument shared network platform.

Key words: deep Web; information extraction; inquiry interface; data mining

0 引 言

科技界有一个流传甚广的故事:一种特别昂贵的大型仪器,在欧洲只有2台,美国只有6台,而在中国已经有几十台了。专家分析,不是因为我们需要那么多,而是因为缺乏共享机制。大型科学仪器投资巨大,使用具有周期性,目前普遍存在信息难以共享、利用率低、重复购置等问题,严重影响科技水平的提高并造成巨大的经济损失。对于大型科学仪器而言,信息的采集和共享显得尤为重要。尤其是在云南省,许多企业面临着科研仪器"吃不饱"和不得不异地求测的难题。因此云南省大型科学仪器协作共用网(简称YSISS系统)的平台建设迫在眉睫,目的是提供仪器共享、优化资源配置并提供管理决策支持。同时仪器的共享必然带来知识的交流和合作,可以说,共享机制会使人们在较近的地方,花较少的钱找到所需要的仪器设备。

云南省大型科学仪器协作共用网以推动云南省大型科学仪器设备资源共享为主要目标,通过建立资源库、信息管理系统等软硬件条件以及制定相配套的运行机制和管理制度,从而有效地提高现有大型科学仪器设备资源的利用率和管理水平,同时也为全国各省市大型科学仪器设备的联合评议工作提供科学、合理、公正的技术支撑,保证新增的大型科学仪器设备资源确实能够避免重复、浪费和布局不合理的弊病,缩短研发周期、省却奔波之苦。

在YSISS系统中,设备管理决策分析子系统需要海量数据的支持,对于这些数据,传统的方法是进行人工采集,费时费力且实时性差。比如:科学仪器的供应商信息,很难人工采集,但实际上,供应商信息在很多大学或企业的网站上已经存在,即隐藏在Deep Web中,但传统的搜索引擎无法检索到。文中主要针对此类问题,设计一种基于Deep Web的信息查询方式,并

收稿日期:2006-12-02
基金项目:国家教育部春晖计划(Z2005-1-53004)
作者简介:王冉冉(1983-),女,江苏沛县人,硕士研究生,研究方向为智能信息系统;黄青松,教授,主要研究方向为智能信息系统。

将所采集的信息整合到本地数据库以便支持管理决策分析。

1 Deep Web的特点

在网络信息资源海量增长的今天,为什么还要研究深网呢?主要原因就是深网具有非常重要的特性,具体来说有以下几个方面:

①相对不可见性,普通搜索引擎无法索引深网。

②资源丰富,数据量大。据Bright Planet公司调查显示,目前存在着超过20万个深网的站点,其资源数量大约为7500TB,是WWW的400~550倍,其中包括5500亿私人文档。

③资源质量相对较高。深网比浅网所涉及的范围更小,内容更为精深,因而质量更高。此外,95%的深网可以免费获取,并且大约一半的深网存在于主体明确的网络数据库中。

④发展迅速。深网是互联网上发展最快的信息资源,从2000年到2004年期间深网已经增长了3~7倍,现已有30.7万个站点,45万个数据库和125.8万个界面[1]。

2 基于Deep Web的查询方式设计

2.1 信息不可见的原因

在大部分的高校或企业网站或数据库中已经载有科学仪器的供应商信息,但它们却无法被传统的搜索引擎检索到,因此可以说这些网络资源对于我们而言是"不可见的"。这些信息属于深网资源,不能被检索到的原因主要有:

(1)未被链接的网页。根据搜索引擎原理,若没有任何其他网页链接指向某一网页,搜索引擎的Spider程序就不能沿着其他网页中的URL爬行到该网页,也就不能将该网页的相关信息搜集到索引库。

(2)网上可检索的数据库。网上可检索的数据库中绝大部分都是结构化的数据。这些数据"隐藏"在网络检索界面后端,存储在Access,Oracle,SQL Server,DB2等数据库系统中。当需要检索数据时,必须使用本网站的搜索工具进行直接查询,在交互式检索窗体中输入检索提问式或选择检索选项,数据库响应请求后,将相应的检索结果按一定的排序规则显示在网页上。网上可检索的数据库可以分为两种类型:可自由获取的公共数据库和需订阅或者付费的数据库[2]。由于搜索引擎的Spider程序尚不具备在交互式检索窗体中填写或选择所需字段信息的能力,无法向数据库提交检索提问式。同时,对于一些必须注册或者付费的网站中的数据库来说,搜索引擎的Spider程序同样没有足够的智能注册后登录系统。因此,无论是哪种类型的数据库,搜索引擎都无法获取其中的数据。

2.2 常见的查询方式

一般来说,搜索深网信息可以从目录指南、具有检索功能的网站、免费数据库,以及专用搜索引擎和优秀普通搜索引擎四种途径入手[3],选择使用相应的检索工具。常见的查询方式主要有:专业目录;专业搜索引擎;主题明确的数据库等。

2.3 基于Web的查询方式的设计

有价值的网络信息一般都存储在数据库中,网上可检索的数据库是深网最大的组成部分,也是深网信息规模大、质量高的最主要原因[4]。文中主要针对网上可检索的数据库进行基于Web的查询方式的设计。

YSISS系统所需要的信息大都存储在此类数据库中,这里针对YSISS系统中的一个实际信息(如表1所示)进行从Deep Web中提取信息方式的设计。

表1 生产厂商信息表

生产厂商名称		供应商类别	
单位联系人		主要仪器产品及性能描述	
营业执照编号		资质类型	
注册资金		营业执照有效期	
提供仪器列表		注册级别	

此类信息很难被传统的搜索引擎所检索到。因为对Spider来说,现在遇到的最大障碍是各种表单。大量的深网资源都是因为Spider不会填写表单而出现的。目前搜索引擎Spider的智能性很低,它的工作方式很简单,只会顺着链接机械地搜集资源。即使遇到很简单的表单,它都无能为力,只能停止搜索,造成了大量深网的产生。若能制造出可以自动填表的机器人,则大量的深网资源就可以浮出水面,为广大用户服务。目前,几个旨在创造更智能化搜索引擎Spider的计划正在实施中,这种Spider能够自动填表和检索信息。该Spider采用两种最基本的办法:一种办法是采用事先设置好的代理程序与特定数据库的表单进行交互;另一种是利用人工智能技术猜测表单所需内容,使Spider能够透过表单进入数据库内部检索信息。这些技术若能够被搜索引擎广泛使用,则会大大改善现状。

现在我们实验小组正在试图设计一套能够自动填写表单的深网查询系统DWQS(Deep Web Query System),建立了表单数据模型,并试图设计一套DWQSL查询语言,能够利用复杂多变的客户端脚本程序。DWQS系统由以下几个部件组成:用户接口、网络文档加载器、HTML剖析器、查询处理器、提取模块和存储/检索管理器、体系结构如图1所示。

此设计思想首先是由查询解析器对用户输入的查询信息,如关键字等进行分析,然后由HTML剖析器

对搜索的网页进行解析,通过表单提取器、标签提取器及 JS 函数提取器对表单信息进行提取。之后将所提取的信息与表单数据库中的已有表单模式进行匹配,若匹配成功,则发出信号,由查询赋值模块对用户信息进行相应的加工赋值,并填写表单,继而传送到网络文档加载器进行搜索,若无相同的表单匹配模式,则发出信号,由值检验器/客户端脚本分析器对所提取的表单信息进行分析,并尝试赋值填表。通过 HTTP 请求模块对填表结果进行检验。若成功,则传送到网络文档加载器进行搜索,同时将此种新的表单模式及赋值填表形式存入数据库以便下次进行表单模式匹配。

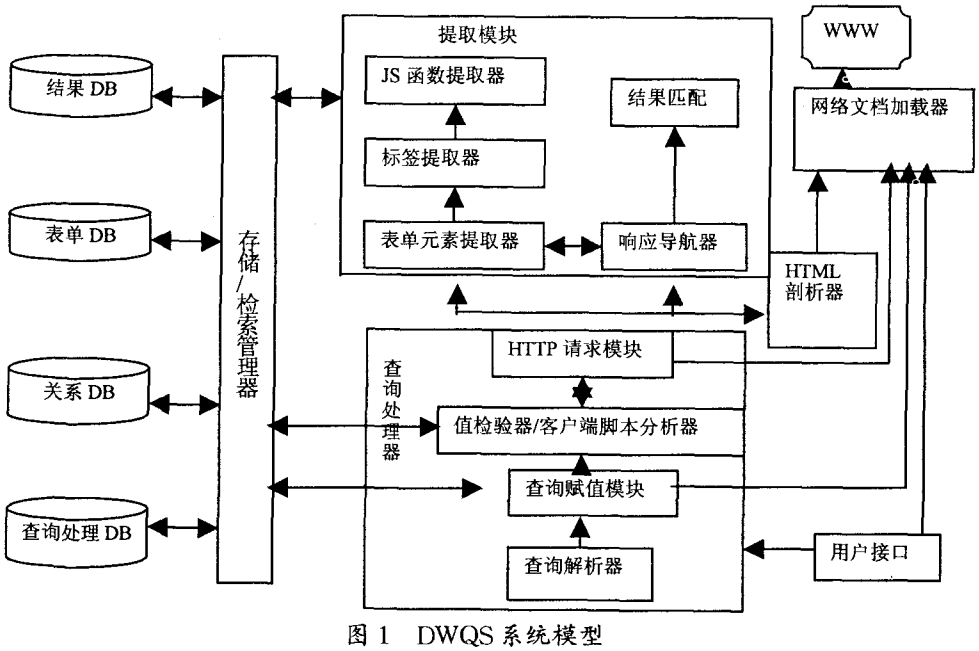

图 1　DWQS 系统模型

其中,提取模块主要是对表单信息进行提取、分析,然后将其与数据库中已经存有的表单模式进行匹配,若匹配成功,则发出信号,由查询处理器进行填表;若匹配不成功,则将表单信息发送至值检验器/客户端脚本分析器,由该模块进行尝试赋值填表。

查询处理器的功能主要是首先对用户输入的查询信息进行分析,然后根据表单提取模块的信号进行相应的处理,若有相同模式的表单,则直接通过查询赋值模块根据用户信息对表单进行填写,然后进行检索;若无相同匹配模式,则通过值检验器/客户端脚本检验器根据表单信息及用户信息进行尝试赋值填表。

存储/检索管理器是数据库与表单提取模块、查询处理器之间的通信转换模块,提取器和处理器可以通过它从数据库中提取表单模式及赋值填表形式,也可以将新的表单模式存入相应的数据库。

2.4　对所得信息进行数据挖掘处理

通过 DWQS 系统所提取的信息并非完全是大型科学仪器的信息,还可能混合了其它信息于其中,因此,这里需要对所得信息进一步进行处理,即数据挖掘处理[5]。其基本过程和主要步骤如图 2 所示。

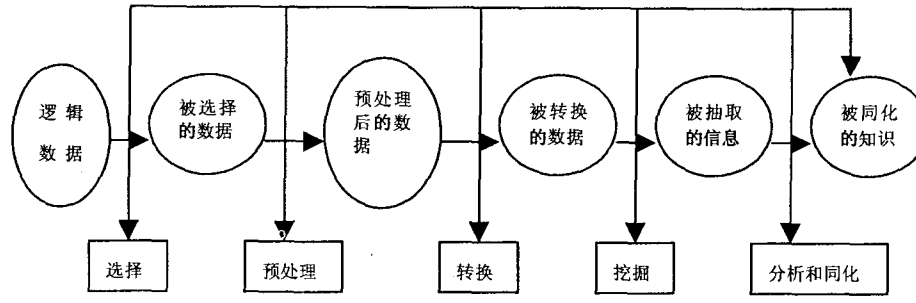

图 2　数据挖掘的基本过程和主要步骤

在对数据进行进一步挖掘处理时,数据的选择主要是搜索所有与业务对象有关的内部和外部数据信息,并从中选择出适用于数据挖掘应用的数据;数据的预处理则是研究数据的质量,为进一步的分析作准备,并确定将要进行的挖掘操作的类型;数据的转换则是将数据转换成一个分析模型,在这里是指表单模型,这个分析模型是针对挖掘算法建立的;数据挖掘是对所得到的经过转换的数据进行挖掘。除了完善选择合适的挖掘算法外,其余一切工作都能自动地完成;结果分析是解释并评估结果,其使用的分析方法一般应按数据挖掘操作而定,通常会用到可视化技术。知识的同化是将分析所得到的知识集成到业务信息系统的组织结构中去。

3　结束语

通过对基于 Deep Web 的查询方式设计以及后期的数据挖掘处理,正试图实现对传统搜索引擎所检索不到的信息进行提取、加工处理。基于 Deep Web 的信息采集系统的优点是将 Deep Web 和数据挖掘的相关

(下转第 177 页)

性,可以通过对同一行业内部有关业务数据描述达成共识,进而设立行业数据标准作为企业之间相关业务数据的参照系。这样各企业就可以依据这个参照系进行数据交换,这就是行业数据交换的理论依据。

下面以企业 A、B 之间的协同服务交互过程为例来说明行业数据交换的原理:

(1)把行业数据标准 X 存储在行业供应链管理平台成为行业标准数据格式模板;

(2)接收 A、B 的企业私有数据格式 F(A)、F(B),存储在企业私有数据格式模板,并建立与行业标准数据格式模板之间的映射关系,即 F(A→X),F(X→A),F(B→X),F(X→B);

(3)依据上述映射关系,由 F(A→X),F(X→B)自动推导出 F(A→B),由 F(B→X),F(X→A)自动推导出 F(B→A),就建立了企业 A、B 间的直接数据交换关系 F(A→B)和 F(B→A),并存储到数据模板映射中心;

(4)企业 A 向企业 B 进行服务请求时,首先检查本地有没有 F(A→B),若没有则到行业供应链管理平台去下载并缓存到本地,然后进行数据交换,即:DATA(A)→DATA(B),这样就得到了 DATA(B),再封装成 SOAP 请求发送到企业 B;

(5)企业 B 接收并解析 SOAP 请求,调用 Web 服务实现处理请求,得到处理结果,检查本地有没有 F(B→A),若没有则到行业供应链管理平台去下载并缓存到本地,然后进行数据交换,即:DATA(B)→DATA(A),这样就得到了 DATA(B),再封装成 SOAP 应答发送到企业 A;

(6)企业 A 接收并解析 SOAP 应答,得到结果数据。整个服务请求、应答过程结束,企业 B 对企业 A 的服务请求同理。

4　结　论

为屏蔽企业计算环境的分布性、异构性,解决行业供应链上企业间应用集成的难题,提出了一个面向行业供应链的企业应用集成架构参考模型,分析了模型实现的关键技术。该模型综合运用了 Web 服务、J2EE 平台和行业数据交换原理,具有简单、开放、安全、高效、标准、可扩展的特点。

文中所提出的集成架构模型已经初步应用于中国摩托车商务平台-协同供应链管理系统,是一个典型的行业供应链的企业间集成的应用。此外,面向行业供应链的企业应用集成架构模型还适用于多种行业(特别是制造行业)供应链上企业之间的集成与协同商务应用,为行业供应链管理系统的开发和实施提供了一个通用的参考模型。

参考文献:

[1] 黄国青,章　勇. 面向供应链管理的企业应用集成技术选择模型[J]. 计算机工程与应用, 2005(23):221-229.

[2] 陈兵兵. SCM 供应链管理——策略、技术与实务[M]. 北京:电子工业出版社, 2004: 626-627.

[3] 陈传波, 张道杰, 李　涛. 基于 Web 服务的企业应用集成模型研究[J]. 计算机工程与科学, 2004,26(12):15-29.

[4] 柴晓路. 技术刨析:传统应用与 Web 服务的接口[EB/OL]. 2002-09-26. http://industry.ccidnet.com/art/732/20020926/26335_4.html.

[5] 张玉东, 刘广钟. 基于 J2EE 平台和 Web 服务的企业应用集成方案[J]. 计算机工程与设计, 2004,25(11):2015-2017.

[6] 殷　庆, 刘卫宁. 面向行业数据交换中间件 EasySwitch 的系统设计与实现[J]. 计算机科学, 2004,31(7):159-162.

(上接第 170 页)

代计算机(专业版),2006(5):93-97.

[2] 毛德操,胡希明. Linux 内核源代码情景分析(上册)[M]. 杭州:浙江大学出版社,2001.

[3] 杨　伟,刘　强,顾　新. Linux 下的存储管理[J]. 电子科技,2005(9):7-10.

[4] Gorman M. 深入理解 Linux 虚拟内存管理[M]. 白　洛,刘森林等译. 北京航空航天大学出版社,2006.

[5] Bryant R E, O'Hallaron D. 深入理解计算机系统[M]. 龚奕利,雷迎春译. 北京:中国电力出版社,2005.02.

(上接第 173 页)

知识相结合,避免了传统手工收集表单信息的弊端,为实现信息提取自动化提供平台。由于技术等条件的限制,本系统仍在设计中。

参考文献:

[1] 张笈秋. 深网的概念、规模及内容[J]. 中国信息导报,2004(10):57—60.

[2] Sherman C, Price G. The Invisible Web : Uncovering Sources Search Engines Can't See[J]. Library Trends,2003(2):282-298.

[3] Lackie R J. Those Dark Hiding Places: The Invisible Web Revealed[EB/OL]. 2005-02-25. http://library.rider.edu/scholarly/rlackie/Invisible/Inv-Web-Main.html.

[4] 吴志强,严贝妮. 从隐蔽网络到国际互联网信息资源控制计划[J]. 图书情报工作,2004,48(3):82-85.

[5] 罗海蛟,刘　显. 数据挖掘中分类算法的研究及其应用[J]. 微机发展,2003,13(5):48-50.