Deep Web 技术在科学数据共享平台中的应用*

郭少杰　陈雅冰

摘　要：Deep Web 中蕴含了海量的可供访问的信息，并且还在迅速地增长。随着互联网应用的发展，网上的在线数据库大量涌现，Deep Web 数据集成成为当前信息领域的一个研究热点。为了方便用户查询数据，对 Deep Web 技术的应用进行了研究，提出了 Deep Web 技术在科学数据共享平台中的架构，并阐述了具体的实现。

关键词：科学数据；科学数据共享平台；Deep Web

0 引言

科学数据库共享平台是将科学数据资源集成并共享，提供科学数据、科技文献、专利、标准等科技资源信息的查询、检索、交流，为广大科技工作者服务、为科技创新服务的平台。该平台使科学数据得到更大泛围的共享、应用和交流，科学数据发挥更大的价值。

科学数据共享平台的科技文献检索、专利检索、标准检索等功能的实现，应用了 Deep Web 技术。

1 Deep Web 技术概述

1.1 什么是 Deep Web

Web 信息按照“深度”划分，可以被划分为 Surface Web 和 Deep Web。其中，Surface Web 是指能被传统搜索引擎检索到的页面，如静态的 HTML；而 Deep Web 则是传统搜索引擎不能检索到的一些内容，主要是指需要用户填写提交一个 HTML 的 Form 表单后才能搜索到的内容。

1.2 Deep Web 搜索过程

搜索 Deep Web 的过程如下（如图 1）：(a) 从网页上获取表单；(b) 对表单进行关键字抽取并集成；(c) 填充表单并提交；(d) 分析返回的结果。

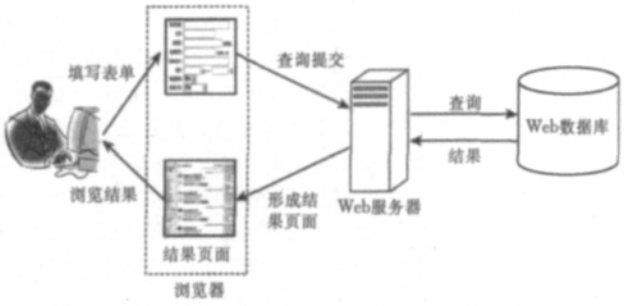

图 1　Deep Web 搜索过程

2 Deep Web 在科学数据共享平台中实现的功能

科学数据共享平台，主要实现了科学数据、科技文献、专利、标准等科技资源信息的查询、检索、交流，其中科技文献检索、专利检索、标准检索等功能的实现，应用了 Deep Web 技术。

2.1 科技文献检索

科技文献是重要的信息资源，包括学术论文、期刊、会议、成果、法规等各类电子资源。应用 Deep Web 技术，可以便捷、快速地获取科技文献，为科技文献有需求者，尤其是一些科研工作者提供帮助与便利。

科技文献模块主要包括科技文献检索和科技文献后台管理两个部分：

(1) 科技文献检索：科技文献查询服务方式有两种：统一查询和高级查询。统一查询主要是按照关键字统一跨库查询；而高级查询可以按照文献的标题、作者、摘要、关键词进行查询。

(2) 科技文献后台管理：科技文献后台管理主要是平台管理员对科技文献数据源参数的设置。数据源参数的设置，主要是对 Deep Web 数据源站点的参数配置，如站点的新增，修改，删除，禁用，开启等功能。

2.2 专利检索

专利检索模块主要分为专利检索、专利检索站点管理。

(1) 专利检索：使用 Deep Web 技术进行检索是把用户在统一接口中输入的查询转发到多个选定的数据源接口表单上以形成目标查询。用户的查询提交后，再获取所有的结果页面，并对结果页面进行信息抽取，然后对所有数据源的检索结果记录进行合并和筛选。最终将满足用户查询条件的结果以统一的格式呈现给用户。

(2) 专利检索站点管理：专利检索站点管理提供了对可进行 Deep Web 搜索的数据源站点列表的管理功能，由管理员选择是否发布，供用户搜索。

2.3 标准检索

标准检索模块分为标准检索和标准检索站点管理两部分。

(1) 标准检索

标准检索的方式分为两类：统一检索和高级检索。

统一检索：用户输入单个关键字，选择关键字的类别，然后勾选所要搜索的数据源站点进行检索。关键字可选择的类别包括名称、标准号等。

高级检索：用户填写一个以上的关键字，勾选所要搜索的网站进行查询。在标准检索的高级查询中，可输入字段包括：标准号、中文标题、英文标题、发布单位、起草单位、中标分类号、

* 基金项目：广东省科技计划项目“广东省科学数据共享平台的研发”（2009B060100042）

国际分类号等。

⑵ 标准检索站点管理

标准检索站点管理提供了对可进行 Deep Web 搜索的数据源站点列表的管理，由管理员选择是否发布，供用户搜索。

3 Deep Web 在科学数据共享平台的实现方法

Deep Web 的框架可以分为三个主要的模块：查询接口集成模块、查询处理模块和查询结果处理模块。

⑴ 查询接口集成模块：为用户提供统一的查询接口，可以同时向多个查询接口提交查询。这一模块主要解决 Web 数据库的发现、查询接口模式的抽取、Web 数据库的分类，查询接口集成等问题。

⑵ 查询处理模块：将用户在集成的查询接口上填写的查询转化到对各个 Web 数据库本地查询接口的查询。这一部分处理 Web 数据库选择、查询转换和查询提交。

⑶ 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。

科学数据共享平台中的 Deep Web 查询模块是基于配置文件的集成系统，它的主要思想是：用统一的集成程序，利用针对每一个网站的配置文件，对 Deep Web 的数据进行集成。集成程序是统一的，而针对网站只需要写配置文件。所使用的框架如图 2 所示。

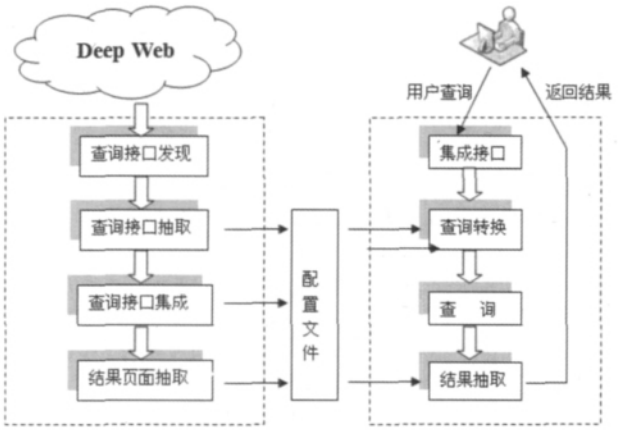

图 2 科学数据共享平台中 Deep Web 搜索引擎框架

首先，管理员对站点进行配置，程序根据管理员的输入生成该站点的查询接口配置文件、结果页面配置文件，以及集成接口配置文件。然后，用户在前台页面输入查询信息，选取需要查询的站点，提交查询，则系统读取相应的配置文件进行处理，然后返回各个站点的查询结果。

3.1 查询接口的发现

为了简化问题，数据源的发现和选择，采用人工查找的方式，避免使用网络爬虫花费大量时间爬取网页。管理员只需要提供查询接口的地址，程序就能够自动生成查询接口配置文件和集成接口配置文件。

网页表单是使用<form>标签标识的。虽然数据源的发现和选择是通过人工查找的，但是网页上仍然存在很多表单并非 Web 数据库的查询接口，例如登陆表单，站内搜索表单等。在这一模块中，我们定义了一些启发式规则，用于判断表单是否为正确的查询接口：

<1><form>标签包含的 HTML 代码中，没有任何内容的，或者一个文本框都没有的表单排除。

<2>关键字过滤。如果表单中的文字，表单元素的属性包含例如“注册”、“登陆”、“密码”等关键字，或者表单属性中包含 google、baidu 等搜索网站的网址的，则排除该表单。

3.2 查询接口的抽取与集成

查询接口模式抽取主要是查询接口属性和控件约束元素的获取与分析，把它们按照逻辑关系重组成一个个属性添加到接口模式集合中。为了让用户更容易理解和使用查询接口，设计者通常会融入多种类型的视觉特征，主要包括位置、布局和外貌等特征。例如基于 Dom 树抽取，基于视觉的网页分割算法等进行模式抽取。

本系统采用基于 Dom 树的方式进行抽取。一个表单的元素通常包括隐藏文本框 hidden、文本框 text、复选框 checkbox、单选框 radiobox、下拉框 select、按钮 button。基本上，每个表单元素的附近都有文字标签表示这些元素的意义。查询接口的抽取就是获取这些元素，并且从表单中分析、匹配，获得每个元素的标签。本系统中，抽取的结果将以 xml 格式写成站点配置文件。

通过大量观察，我们使用一系列启发式规则来抽取查询接口。根据这些规则，我们记录了表单元素的属性及其各自的标签，生成 xml 格式的文件作为查询接口的配置文件。配置文件记录了站点资源和和表单 form 的信息。表单信息中，包含了 6 个组，分别记录了隐藏文本框 hidden、文本框 text、复选框 checkbox、单选框 radiobox、下拉框 select、按钮 button 的情况，每组信息都记录了元素的基本属性以及程序分析得到的标签。

集成配置文件是根据各个查询接口配置文件生成，采用了局部集成方式。集成配置文件的生成需要一个标签匹配的过程。匹配分精确匹配，模糊匹配两轮。第一轮是精确匹配，即字符串完全相等的匹配。成功匹配的将不再参加第二轮的匹配。在模糊匹配中，如果表单元素标签含有全局名称的所有字符，并且相对顺序一致，则认为是成功匹配。

3.3 查询的提交

查询的提交模拟了用户通过浏览器访问网站，填写表单提交查询的过程。查询提交的过程如下：⑴ 访问查询接口网站，获得网站的头部信息。因为实际应用中，存在一部分网站通过头部信息记录用户查询信息，缺少这些信息就会导致不能正确传递查询条件。⑵ 填写提交的参数。对于文本框元素，根据集成接口配置文件填写参数，对于下拉框、单选框、隐藏文本框则使用网页本身的默认值，对于复选框则全选所有的值。⑶ 提交查询，获取返回结果。如果网站返回跳转码，则继续跟踪访问，直到返回成功码或错误码停止。用户查询界面如图 3 所示。

3.4 结果页面的抽取

结果页面的抽取首先需要分解网页中有用的信息单元和无效的信息单元，也即网页清洗技术，从 Web 页面中划分出精确的信息单位，并根据 Web 页面信息加工的后续应用需求，将网页中不需要的部分剔除掉，同时抽取出有用的信息。

网页清洗主要有三步：首先是去除页面中的注释、脚本、样式等无关信息，然后再将页面或为若干块，大致可以分为文本数据块、链接块、图像块。最后按照语义对各块作进一步的区

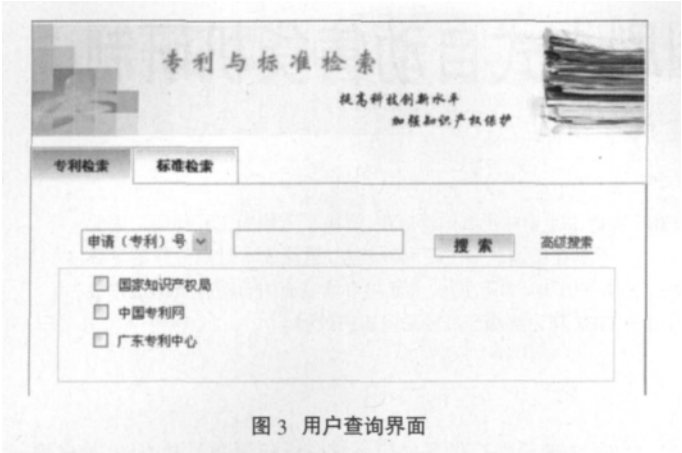

图 3 用户查询界面

分，从文本块中分出广告等杂质信息块，从链接块中分出相关链接块、导航链接块等。经过上述处理后，Web 页面在结构和语义上都被划为细粒度的信息块，然后就可以对这些数据进行后续的信息加工处理，抽取出更加详细的数据信息。

基于 Web 的信息抽取，大致有三种，一种是基于 html Dom 树结构，一种是基于人工智能的语义抽取，一种是基于规则定义方式。Dom 树方式，虽然很好的显示数据相互之间的依赖关系，但是在数据抽取上效率不是很好，它需要将 html 标签解析为 Dom 树，然后对其进行树结构操作，尤其是对树结构的遍历上比较耗费资源。人工智能中语义抽取方法，是利用领域中数据的元数据描述信息进行语义判断，虽然可以部分实现自动化抽取数据的效果，但需要建立大量的语义库，而语义库的建立是现在人工智能领域中一项棘手问题。基于规则的抽取方法，是利用网页中数据的固定显示格式，配置抽取规则，然后对网页数据进行抽取，这种方式，配置上也不是很复杂，只要指定好抽取规则，可以快速的抽取到数据信息。

基于领域的 DeepWeb 搜索引擎是针对一些特定的网络资源，一些特定的网站进行数据抽取，在特定的网络资源中，数据显示格式呈现出一种结构化的、有规律的样式。如文献资源，常见的数据项有标题、作者、出处、发表时间、摘要等相关信息。而这些网站数据在特定的网站中结构上相当稳定，基本上都是以列表的形式呈现给用户。而这些数据在 html 源码中的格式基本一致，其数据的格式基本上可以按照数据域、数据块、数据项分。数据域是包括要抽取信息的最小子集（在 html 结构中是以 html 标签的形式），数据域中包括多个数据块（即数据记录），数据项是抽取的最小字段（即要抽取出的数据）。

在该系统的实现中，首先将 web 数据划分为三种数据区：数据域、数据块、数据项。其中数据域包括多个数据块，数据块包括多个数据项，数据项就是我们要抽取的数据信息。

抽取信息结构的划分是根据 html 结构，根据页面的 Dom 树结构和正则表达式相结合方法抽取出特定数据区。通过指定数据域、数据块、数据项的首尾标识信息（这是配置的内容，需要查看网页的源码，也可以部分的做成自动匹配，但是匹配的准确度不是很高），然后再利用正则表达式进行准确、快速抽取出数据信息。只要准确配置，只要是页面上出现的数据都能抽取出来。同时，页面平台也是实现了翻页查找数据的功能，通过正则表达式抽取出网页中下一页的链接信息，可以自动模拟表单提交，检索网页中多页数据，同时对多页数据进行抽取。

多页数据的处理是利用正则表达式抽取出网页中所有链接，然后对链接的 URL 进行比对，判断其 value 的属性值，然后抽取出其 url，继续后续同样的抽取工作。平台实现后的效果如图 4 所示。

图 4 查询结果列表

4 结束语

我们所设计的系统为用户提供了统一的访问接口，使得用户可以方便地进行科技文献、专利信息、标准信息的查询。将本系统进行适当的参数调整，也可以应用于其他领域。Deep Web 技术在科学数据共享平台中的应用是一次成功而且有意义的尝试。Deep Web 数据集成还有巨大的研究空间，仍需不断进行探索。

参考文献：

[1] 刘伟，孟小峰，孟卫一. Deep Web 数据集成研究综述 [J].计算机学报，2007，30（9）：1475-1489.

[2] 钟昕，伏玉琛. 书籍搜索领域 DeepWeb 数据集成系统 [J]. 计算机技术与发展，2008 18（9）

[3] Cai D，Yu S，Wen J，at al. Extracting Content Structure for Web Pages Based on Visual Representation [C] In AP Web，2003，Xi'an，2003:406-417

[4] Cai D，Yu S，Wen J，at al. VIPS:a Vision-based Page Segmentation Algorithm [R] Microsoft Research Technical Report，MSR-TR-2003-79，2003.

[5] 李朝，彭宏叶，苏南，张欢，杨亲遥. 基于 DOM 树的可适应性 Web 信息抽取. [J] 计算机科学，2009 36（7）

[6] 廉成洋，毛宇光，黄玉明，基于启发式规则的 Web 信息抽取技术研究. [J] 计算机技术与发展，2009 19（8）

（作者单位 郭少杰，广东省科技信息中心 陈雅冰，华南理工大学计算机学院）