ISSN 1000-0054 清华大学学报（自然科学版）2005 年 第 45 卷 第 S1 期 36/39
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2005, Vol. 45, No. S1 1896-1902

Deep Web 爬虫研究与设计

郑冬冬， 赵朋朋， 崔志明

（苏州大学 计算机科学与技术系，苏州 215006）

摘 要：随着 Web 的发展，越来越多的数据可以通过表单提交来获取，这些表单提交所产生信息是由 Deep Web 后台数据库动态产生的。在这种情况下，信息集成就更加需要 Web 爬虫来自动获取这些页面以进一步地处理数据。为了帮助用户完成这样的任务，提出一种用于搜集 Deep Web 页面的爬虫的设计方法。此方法使用一个预定义的领域本体知识库来识别这些页面的内容，同时利用一些来自 Web 站点的导航模式来识别自动填写表单时所需进行的路径导航。通过对来自不同领域的 Deep Web 站点的大量实验，验证了此方法是非常有效的。

关键词：Deep Web；导航模式；领域本体知识库；爬虫

中图分类号：TP 393 **文献标识码**：A

文章编号：1000-0054(2005)S1-1896-07

On the research and design of deep web crawler

ZHENG Dongdong, ZHAO Pengpeng, CUI Zhiming

(Department of Computer Science and Technology, Soochou University, Suzhou 215006, China)

Abstract: As the web grows, more and more data has become available under dynamic forms of publication, such as legacy databases accessed by an HTML form. In this way, integration of this data relies more and more on the Web Crawler that can automatically fetch pages for further processing. As a result, there is an increasing need for tools that can help users generate such agents. The method is described for automatically generating agents to collect Deep Web pages. This method uses a pre-defined ontology repository for identifying the contents of these pages and takes the advantage of some patterns that can be found among web sites to identify the navigation paths to follow. The results of a number of experiments carried out with sites from different domains demonstrate the accuracy of the method.

Key words: deep web; navigation patterns; domain ontology repository; crawler

Deep Web，Hidden Web 和 Invisible Web 均指同一个概念，它是与 Surface Web 相对应的概念，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 和 Gary Price 对 Deep Web 定义为：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、文件或其他通常是高质量、权威的信息。最近对 Deep Web 的研究[1]得到了一些有意义的发现：

1）目前 Deep Web 大约有 307 000 个站点，450 000 个后台数据库和 1 258 000 个查询接口。其信息资源仍在迅速增长，从 2000 年到 2004 年，它增长了 3～7 倍。

2）Deep Web 内容分布于多种不同的主题领域，尽管电子商务是主要的驱动力量，Web 数据库的发展趋势不仅在此领域，同时在非商业领域占的比重相对更大些。

3）当今的 Web 爬虫并非完全爬行不到 Deep Web 后台数据库内容，当前主要的搜索引擎已经覆盖 Deep Web 大约 1/3 的页面。然而，在 Deep Web 信息覆盖率上当前搜索引擎存在技术上的本质缺陷。

4）Deep Web 后台数据库由结构化的和非结构化的组成，其中结构化的是非结构化的 3.4 倍之多。

5）虽然一些 Deep Web 目录服务已经开始索引 Web 数据库，但是其覆盖率比较小，仅为 0.2%～15.6%。

6）Web 数据库往往位于站点浅层，94%之多的大量 Web 数据库可以在站点前 3 层发现。

由此可以看出 Deep Web 所含信息量要比 Surface Web 信息量多的多，Deep Web 后台数据库多是结构化的关系数据库，因此信息的质量要比

收稿日期：2005-05-20

基金项目：Deep Web 关键技术研究

作者简介：郑冬冬(1980-)，男(汉)，河南，硕士研究生。

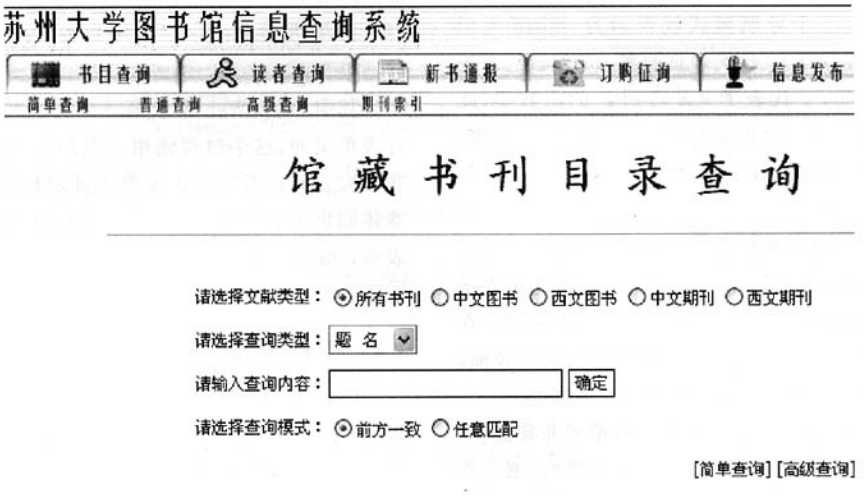
电话 0512-65112730 转 802

E-mail: udbtuu2003@163.com

非结构化的数据要高。然而不能直接使用数据库技术管理和查询这些数据。处理这些数据的一种方法是使用 Web 包装器[2,3]，它是一种可以从 Deep Web 中抽取数据然后以 XML 或关系表的形式存储起来的程序。

传统的爬虫仅能爬行所谓的公共可索引的页面，它是通过分析页面中超链接来爬行的。然而 Deep Web 页面是为响应来自客户端的表单查询请求由服务器端后台数据库动态产生的。因此要获取这个页面集，就需要一种特殊的爬虫来搜集这些页面。

为了解释 Deep Web 爬虫的功能，假设想获取来自苏州大学图书馆站点的含有丰富信息的页面集。首先进入苏州大学图书馆主要查询页面，如图 1a 示。其中有一个基于关键字的简单搜索框。同时用户可以使用高级搜索表单，如图 1b 示。其提供的多个属性来满足更高要求的搜索任务。当用户在表单中填入相应字段后，将会返回一个结果页面。结果页面可能由多页索引组成，可以通过点击“下一页”等类似按钮操作来获得其他结果页面。



(a) 简单搜索页面

(b) 高级搜索页面

图 1 苏州大学图书馆主页查询

本文提出了一种用普通的导航特征描述的Deep Web爬虫的设计方法。此方法使用了启发式函数集和一个领域本体知识库来自动发现相关表单,同时填写表单和搜集含有匹配结果的页面集。并通过对来自不同领域的Deep Web站点的大量实验,证明了所述方法是非常有效的。

1 导航模式

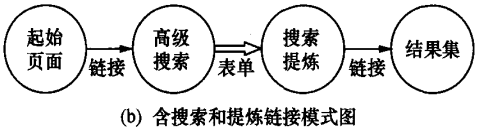
产生动态页面的方法多种多样,基于此因素仅考虑两种通过用户在Web站点上的导航行为归纳出来的通用的导航模式。其形式化的定义如下:

定义 1: 一个导航模式被表示为一个五元组 $\Re=(P,\Sigma,\sigma,p_0,T)$, P 代表页面集, Σ 代表页面间的转换机制集, σ 代表 $P\times\Sigma$ 转到 P 的函数集,此函数表示用户从一个页面转到另一页面的行为集, $p_0\in P$ 是起始页面,且 T 是 P 的子集,代表用户访问的相关页面集。

用一个有向图来代表导航模式,其中图中每个元素是 P 中的一个点,顶点 p_i 与 p_j 之间的边表示通过转化机制得到的 p_i 与 p_j 页面的一条路径。在Web站点上用户操作的每一个特定序列称为导航。每一个导航序列代表导航模式中的一个实例。

在Web站点设计中有两种导航模式非常流行。第一种如图2a所示,用户从站点主页开始,填写搜索表单,然后提交表单获取结果页面集。第二种增加了含有用来搜索提炼链接结果集的中介页面,从此页面用户可以获取最终结果页面,如图2b示。在此设计的Deep Web爬虫将仅限制于上述的两种导航模式。

(a) 一般导航模式图



(b) 含搜索和提炼链接模式图

图2 导航模式图

2 Deep Web爬虫

为了实现自动获取Deep Web页面的任务,爬虫必须能模拟用户在各个站点上的导航,即此爬虫必须能跟踪超链接,填写表单,最后获取和识别结果页面。实现这样的爬虫,传统的方法是用特定的语言代码来实现,此方法非常耗时同时常伴随有错误发生,维护起来代价也比较大。另一种方法是使用ASByE[4]工具,其中用户提供导航实例用于完成代理方案。然而这种方法始终都要求代理工作过程中有用户的参与,这就导致在多个站点迭代执行的过程冗长,代价也相应增加。

所设计的爬虫能针对特定应用领域填写表单和取回所有返回结果页面。其工作过程可分为3步:

1) 寻找表单; 2) 学习填写表单; 3) 识别和取回结果页面。

其中Deep Web爬虫第一步从站点主页开始爬行表单页面,这个过程使用一组启发式规则来去除非研究表单; 第二步从表单中抽取标签,配合领域本体知识库的参与,爬虫尽力学习如何正确地填写表单; 最后一步提交表单,然后取回结果页面识别记录。整个过程如图3所示。

2.1 领域本体知识库

Deep Web爬虫的一个关键部件是领域本体知识库。它用来识别特定应用领域里的知识,它实质上是关于一个主题知识的对象集。领域本体知识库事实上是表单上的标签属性-值对的集合(OLVS)。〈l,v〉用来描述应用领域对象集里的对象, l表示对象类型, v表示在一定值域内对象的值。

领域本体知识库中对象的值可以来自一个数据库或特定应用程序的输出,甚至可以通过领域专家手工建立一个这样的知识库。而设计这样一个知识库关键是对象的表示方式。

通过抽取网络上用户感兴趣的特定领域资源数据的方法来构建这个领域本体知识库。例如对书籍领域对象的数据抽取[5,6]可以通过从返回页面中抽取使用的关键词"C++"或"Network"等。实验使用DEByE tool[7]工具来进行这些数据的抽取工作。图4展示了书籍对象集内抽取的一个对象实例。其中类型t使用属性类型是ATOM元素,值v用VALUE元素来表示。因此对于这个对象,使用一属性-值对〈Title,"Java程序设计,第二版"〉,〈Author,"李明,等."〉,和其他属性值对〈…,…〉。

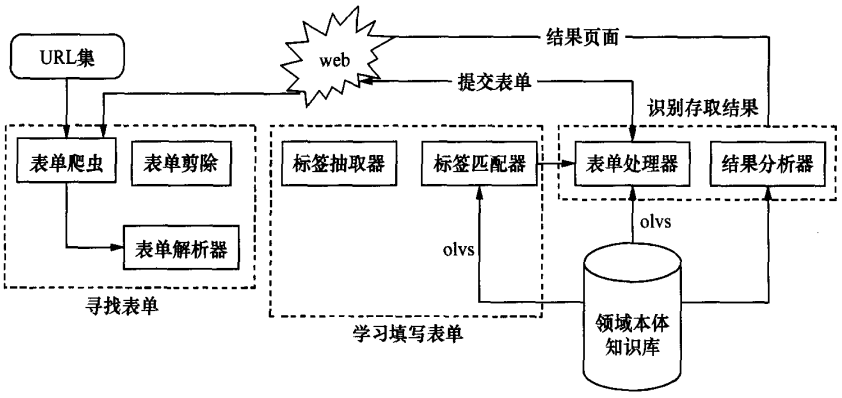


图 3 Deep Web 爬虫系统框架

```
<?xml version = "1.0"><?xml: stylesheet type = "text / xsl"href = "nested.xsl"?>
<OBJECTS>
  <TUPLE type = "SuDaLibrary">
    <ATOM type = "Title'>
      <VALUE><![CDATA[Java程序设计，第二版]]></VALUE>
    </ATOM>
    <ATOM type = "Author">
      <VALUE><![CDATA[李名，等]]></VALUE>
    </ATOM>
    (... ...)
  </TUPLE>
  (... ...)
<OBJECTS>
```

图 4 领域本体知识库中对象实例

2.2 寻找表单

为设计关于特定领域知识 S 的 Deep Web 爬虫，起始于一个 Web 目录，此目录中包含有领域 S 信息的多个站点 URL。这些 URL 被认为是搜索 Deep Web 站点的入口点。

对于每个入口点，首先进行盲目的宽度优先搜索来寻找 HTML 表单，限制搜索深度和取回页面的数量。这一步中主要目标是定位于高级搜索表单或者任何可以返回丰富后台数据库内容的表单页面。实践中可以找到很多这样的表单，它们形成一个候选表单集。然而它们中的很少一部分是相关表单，因此有必要运用一些启发式规则来去除那些不合要求的表单。启发式规则定义如下：

启发式规则 1：给定一个阈值 ξ，具有 $n<\xi$ 的表单将被忽略不予考虑，n 表示 Web 表单中要填写的字段个数。

启发式规则 2：对于给定表单 ζ，如果 ζ 含有密码等的 HTML 类型元素，ζ 表单将被忽略不予考虑。

跟 HiWE 系统[8]类似，规则 1 目标是去掉那些用于一般搜索的简单表单。规则 2 用来去除那些需要进一步注册，购买或要登陆的表单。尽管这两个规则非常简单，但在实验中它们可以非常有效的去除非研究的表单。

然而在一些站点中没有高级搜索表单同时使用这两条启发式规则之后所有候选表单将被全部去除。在此情况下，通常对具有相应文本标签简单搜索框的表单也将视为系统研究的表单。称这样类型的表单为带限制的简单表单(如图 1a 所示)。因此当所有候选表单都被去除的情况下，尽力寻找带限制的简单表单。对于既不是高级搜索表单的又不是带限制的简单表单的都不是要研究的表单。

2.3 学习填写表单

运用以上两条启发性规则去除非法表单后，剩余的表单将是要研究的对象。下面将详细介绍在领域本体知识库的参与下，爬虫如何来学习自动填写

表单[8-10]。表单填写任务包含寻找表单字段和领域本体知识库中对象属性之间的映射关系。解决这个问题的方法来源于两个实践发现：1) 为了使用户任务容易，Web 设计者尽量避免复杂的导航模式；2) 事实上大部分表单字段都有文本标签与之相对应，它们通常非常相似。因此要把领域本体知识库对象属性映射到表单字段的惟一线索就是与表单字段相应的文本标签。然而 HTML 并不提供用来识别标签和相应字段映射关系的明确机制。可以使用另一启发式规则来获取表单字段相应标签，通过确定标签与其邻接的表单字段域位置关系。通常标签位于表单字段域的左边或者上面(如图 1b 所示)。

与 HiWE 系统中的 LVS 表不同，爬虫使用代价相对较低的解析技巧来抽取表单标签。在此使用了一个事件驱动的解析器和两个缓冲区，其中一个用于表单字段上左边的文本标签，另一个用于表单字段上面的文本标签。下面给出这个启发式规则：

启发式规则 3：对于给定表单 ζ，其有两个缓冲区 β_l 和 β_a(一个代表表单字段上左边的文本标签，另一个代表表单字段上面的文本标签)。每一个表单 ζ 中被解析的本文标签首先插入 β_l，当遇到终止符(如：〈BR〉或〈TR〉)时将 β_l 中的内容复制到 β_a 中。当表单 ζ 发现一个输入字段时，通过 $l_f \leftarrow \mathrm{NomEmpty}(\beta_l, \beta_a)$ 函数来委派相应字段 f 的标签，同时把 β_l 和 β_a 清空。

当识别表单标签后，配合领域本体知识库来确认表单是否是要是一个真正相关的表单。通过表单标签和领域本体知识库中对象属性间的映射来寻找相应表单字段的值。如果没有发现任何匹配值，这样的表单也将被抛弃。对于带限制的简单表单，这种方法是比较直接的，因为可以很容易获得单一标签的匹配值，由于其标签值很容易从 HTML 标记〈SELECT〉中找到。如果发现了相应标签的匹配值，爬虫将很容易知道如何填写表单，因为所有必要的提交字段值都可以在领域本体知识库中得到，接下来就是表单的提交。下面就介绍表单提交后的工作。

2.4 识别和获取目标页面

当识别搜索表单集之后，必须模拟爬行站点最优的导航模式。通常要考虑在提交表单和返回丰富结果页面集间是否存在用于提炼结果的中介页面。进一步的，必须识别中介页面和结果页面间的转换机制。

为正确选择爬行站点的导航模式，系统设计了一个测试。这个测试的目的是用来核对某给定的 HTML 页面是否是一个含丰富数据的页面。下面的启发式规则 4 用于测试某给定 HTML 页面中是否有与领域本体知识库中相似的对象，以此来判断给定 HTML 页面是否是含丰富数据的结果页面。

启发式规则 4：设 ξ 是一 HTML 页面，ζ 是具有〈l,v〉结构的对象集构成的领域本体知识库。如果可以识别出页面至少含 n 个对象 O，且有 $O \in \zeta$ 则认为 ξ 是一含有丰富数据的页面(即结果页面)。

为识别这些对象和实现丰富页面的测试，使用 Golgher 等[9]提出的相关技术，它可以用来自动识别 HTML 中一个完整对象是否类似与领域本体知识库中的对象。注意到领域本体知识库中对象与表单对象有很大的相似概率。实验中，如果可以识别某 HTML 页面中至少一个完整的对象在领域本体知识库中，就可以认为该页面是一个含丰富数据的结果页面。也就是说，在使用启发式规则 4 时，系统设置 $n=1$。

对丰富数据页面测试，方法是使用丰富数据页面包装器。丰富数据页面包装器是一个通用网关接口(CGI)或基于一些给定输入参数可以动态产生丰富数据页面的脚本。在导航模式的描述中，丰富数据页面包装器是 CGI 或者通过表单提交操作激活的脚本。正像在图 2b 中所述的导航模式，我们使用丰富数据页面包装器进一步爬行站点，跟踪结果页面中的链接集获取更多表单提交的结果页面。

启发式规则 5：从所有结果页面中可获取的其他链接中，至少要寻找类似“下一页”和“更多”等这样的关键词。然后跟踪链接核对新页面是否是含有丰富数据的页面。若获取了结果页面，包装器保存针对这种链接的模式表达式。当爬虫运行的时候，匹配这种模式表达式的所有页面将会被搜集起来。

通过启发式规则 5 我们就可以获取表单提交后所有含丰富数据的结果页面。为了运行此爬虫，用户还必须提供一些自动填写表单所需要的参数集。运行爬虫后搜集到大量 Deep Web 页面，然后可以对其进行信息抽取，最终可以进一步提供信息检索服务。

3 实验分析

在此系统配合第三方开发的领域本体知识库，其主体由 Java 语言来实现。实验评估了爬虫处理过程每一步的性能，验证了每条启发式规则对爬虫效

能的影响。表 1 选取了来自汽车，书籍，软件每个领域知识里的 10 个 Deep Web 站点。对于每个领域，使用 site. baidu. com 目录里的 10 个站点作为系统的入口点。

从大量可获得的表单中可以看到几乎所有站点都使用了一个简单的搜索框。而它们不是要研究的对象，要研究的表单是带限制的简单表单或高级搜索表单。为评价实验结果，本系统使用了信息检索领域里评价标准：准确率和召回率。在此，准确率被定义为使用一些启发式规则后所剩余的相关表单数占没有使用启发式规则时所剩表单总数的百分比。召回率定义为使用启发式规则后剩余的相关表单数占整个实际相关表单数的百分比。

实验首先评估启发式规则 1 和规则 2 对去除非相关表单的有效性(表 2)。为了验证，通过手工核对了程序搜索过程中搜集到的每个表单。这两个启发式规则不仅用于验证去除非法表单，还用于确定是否有合法表单被去除。经过这两个规则的使用，剩余表单少于 3.4%。实验还可得出是否有合法的表单被去除。

然后来评价填写表单的方法，启发式规则 3 用于消除表单标签与领域本体知识库中对象的属性名不匹配的那些表单。表 3 列出正确处理的相关表单数和去除的不匹配表单数，其准确率和召回率都达到了 100%。

方法最后一步是获取含有丰富数据的结果页面。使用启发式规则 4 可以正确识别是否是含有丰富数据的结果页面。当识别了第一页后，启发式规则 5 用于发现是否还有其他链接能获得结果页面。然而表 4 中的结果是使用启发式规则 5 而获得的平均召回率，它达到了 90%以上。表 5 总结了爬虫整体性能结果。结果表明此爬虫在没有用户参与的情况下，在 3 个领域里信息抽取结果的覆盖率达到了 80%以上。

表 1 实验所使用的站点

汽车	软件	书籍
jsp. auto. sohu. com	www1. skycn. com	www. dangdang. com
price. pcauto. com. cn	www. pconline. com. cn	innopac. lib. tsinghua. edu. cn
auto. sohu. com	download. winzheng. com	lib. cpums. edu. cn
www. qiche. com. cn	dl. 163. com	library. suda. edu. cn
www. chetx. com	down. beareyes. com. cn	www. lib. pku. edu. cn
www. xinhua. org	www. caiqing. net	www. bol. com. cn
auto. china. com	www. k169. net	www. lib. sdu. edu. cn
auto. eastday. com	download. yesky. com	www. 99read. com
auto. tom. com	down1. tech. sina. com. cn	202. 115. 40. 7:8080
www. cheshi. com. cn	www. caiqing. net	www. lib. neu. edu. cn

表 2 表单识别

领域	可获取表单数	实际相关表单数	使用规则 1,2 剩余表单数	剩余表单所占比例/%	召回率/%
汽车	1 443	10	45	3.1	100
软件	1 109	10	38	3.4	100
书籍	967	10	31	3.2	100

表 3 填写表单

领域	相关表单数	使用规则 1,2 剩余表单数	实际匹配表单数	准确率/%	召回率/%
汽车	10	45	10	100	100
软件	10	38	10	100	100
书籍	10	31	9	100	100

表 4 跟踪其他链接

领域	使用多页显示结果站点数	可以跟踪链接的	召回率/%
汽车	10	9	90.0
软件	8	7	87.5
书籍	10	10	100

表 5 整体结果

领域	站点总数	不能填写的表数	不能跟踪链接的	填写的表单数	填写表单所占比例/%
汽车	10	0	1	9	90
软件	10	2	1	7	70
书籍	10	0	1	9	90

4 结论及将来的工作

本文一个主要贡献是提出了一种用于搜集Deep Web页面的爬虫设计方法。此方法可以处理具有复杂结构的站点,同时爬虫工作过程中不需要用户参与。尽管被限制的站点仅有两种导航模式,此方法在不同领域内大量的Deep Web站点上收到了很好的效果。它使用启发式规则集和领域本体知识库来自动发现相关表单,填写表单,同时识别和收集相关结果页面。所提供的启发式规则集实验证明其爬行覆盖率达到了80%以上。

本文另一个重要贡献是使用了领域本体知识库,提出了基于本体的关于特定领域信息抽取系统[11,12]一般框架。由于它是基于本体模型的,这样一来系统可以处理来自任意数据源的页面,因为只要在领域本体知识库中添加相应领域的本体知识就可以扩展其应用范围,从而系统具有很好的扩展性。

本系统在很多方面还可以继续改进以提供更好的服务。如:信息发现阶段使用智能的URL选择器和并行化处理方法将使系统可以扩展更多的数据源;本文对结果中错误消息页面的处理并没有给出处理方法;对于表单标签和领域本体知识库中对象属性匹配算法还有待于改进等,所有这些将是我们下一步的工作。

参考文献 (References)

[1] HE Bin, Patel Mitesh, ZHANG Zhen, et al. Accessing the Deep Web: A Survey [R]. Department of Computer Science, UIUC, 2004.

[2] Laender A H F, Ribeiro-Neto B, Silva A S da, et al. A brief survey of Web data extraction tools [J]. *SIGMOD Record*, 2002, **31**(2): 84-93.

[3] Muslea I, Minton S, Knoblock C. Hierarchical wrapper induction for semistructured information sources [J]. *Autonomous Agents and Multi-Agent Systems*, 2001, **4**(1/2): 93-114.

[4] Golgher P B, Laender A H F, Silva A S da, et al. An example-based environment for wrapper generation [A]. Proceedings of the 2nd International Workshop on The World Wide Web and Conceptual Modeling [C]. USA: Salt Lake City, 2000. 152-164.

[5] Arvind A, Hector G M. Extracting structured data from Web pages [A]. ACM SIGMOD [C]. 2003.

[6] Barbosa L, Freire J. Siphoning hidden-web data through keyword-based interfaces [A]. SBBD [C]. 2004.

[7] Laender A H F, Ribeiro-Neto B, Silva A S da. DEByE—data extraction by example [J]. *Data and Knowledge Engineering*, 2002, **40**(2): 121-154.

[8] Raghavan S, Garcia-Molina H. Crawling the hidden Web [A]. Proceedings of the 27th International Conference on Very Large Data Bases [C]. Italy: Rome, 2001. 129-138.

[9] Liddle S, Embley D, Scott D, et al. Extracting data behind Web forms [A]. Proceedings of the Workshop on Conceptual Modeling Approaches for e-Business [C]. Finland: Tampere, 2002. 38-49.

[10] Modica G, Gal A, Jamil H M. The use of machine-generated ontologies in dynamic information seeking [A]. Proceedings of the 9th International Conference on Cooperative Information Systems [C]. Italy: Trento, 2001. 433-448.

[11] Walker Troy. Automating the extraction of domain-specific information from the web-a case study for the genealogical domain [R]. Brigham Young University, 2004.

[12] CHEN Xueqi. Query Rewriting for Extracting data behind Html Forms [R]. Brigham Young University, Provo Utah, 2004.

Deep Web爬虫研究与设计

作者： 郑冬冬， 赵朋朋， 崔志明， ZHENG Dongdong， ZHAO Pengpeng， CUI Zhiming
作者单位： 苏州大学,计算机科学与技术系,苏州,215006
刊名： 清华大学学报（自然科学版） ISTIC EI PKU
英文刊名： JOURNAL OF TSINGHUA UNIVERSITY(SCIENCE AND TECHNOLOGY)
年，卷(期)： 2005,45(9)
被引用次数： 15次

参考文献(12条)

1. HE Bin;Patel Mitesh;ZHANG Zhen Accessing the Deep Web: A Survey 2004

2. Laender A H F;Ribeiro-Neto B;Silva A S da A brief survey of Web data extraction tools 2002(02)

3. Muslea I;Minton S;Knoblock C Hierarchical wrapper induction for semistructured information sources [外文期刊] 2001(1/2)

4. Golgher P B;Laender A H F;Silva A S da An example-based environment for wrapper generation 2000

5. Arvind A;Hector G M Extracting structured data from Web pages 2003

6. Barbosa L;Freire J Siphoning hidden-web data through keyword-based interfaces 2004

7. Laender A H F;Ribeiro-Neto B;Silva A S da DEByE-data extraction by example 2002(02)

8. Raghavan S;Garcia-Molina H Crawling the hidden Web[外文会议] 2001

9. Liddle S;Embley D;Scott D Extracting data behind Web forms 2002

10. Modica G;Gal A;Jamil H M The use of machine-generated ontologies in dynamic information seeking [外文会议] 2001

11. Walker Troy Automating the extraction of domain-specific information from the web-a case study for the genealogical domain 2004

12. CHEN Xueqi Query Rewriting for Extracting data behind Html Forms 2004

本文读者也读过(8条)

1. 郑冬冬. 崔志明. ZHENG Dong-dong. CUI Zhi-ming Deep Web爬虫爬行策略研究[期刊论文]-计算机工程与设计 2006,27(17)

2. 袁柳. 李战怀. 陈世亮. YUAN Liu. LI Zhan-Huai. CHEN Shi-Liang 基于本体的Deep Web数据标注[期刊论文]-软件学报2008,19(2)

3. 黄聪会. 张水平. 胡洋. HUANG Cong-hui. ZHANG Shui-ping. HU Yang 主题Deep Web爬虫框架研究[期刊论文]-计算机工程与设计2010,31(5)

4. 曾伟辉. 李淼. 曾伟辉 深层网络爬虫研究综述[期刊论文]-计算机系统应用2008,17(5)

5. 周立柱. 林玲. ZHOU Li-Zhu. LIN Ling 聚焦爬虫技术研究综述[期刊论文]-计算机应用2005,25(9)

6. 马安香. 张斌. 高克宁. 齐鹏. 张引. Ma Anxiang. Zhang Bin. Gao Kening. Qi Peng. Zhang Yin 基于结果模式的Deep Web数据抽取[期刊论文]-计算机研究与发展2009,46(2)

7. 冯明远. 林怀忠. FENG Ming-yuan. LIN Huai-zhong 基于最优查询的多领域deep Web爬虫[期刊论文]-计算机应用研究2009,26(9)

8. 汪涛. 樊孝忠 主题爬虫的设计与实现[期刊论文]-计算机应用2004,24(z1)

引证文献(15条)

1. 孟敬. 刘寿强 基于Deep Web Search技术的主题式爬虫模块研究与设计[期刊论文]-科技导报 2011(21)

2. 兰洋. 尤磊 Deep Web中基于关联规则的整体模式匹配[期刊论文]-信阳师范学院学报（自然科学版） 2009(4)

3. 苏晓珂. 赵磊. 黄青松 Deep Web中基于迭代的查询方式[期刊论文]-云南民族大学学报（自然科学版） 2007(1)

4. 辛洁. 崔志明. 赵朋朋. 张广铭. 鲜学丰 基于MapReduce虚拟机的Deep Web数据源发现方法[期刊论文]-通信学报 2011(7)

5. 周二虎. 张水平. 胡洋 基于Deep Web检索的查询结果处理技术的应用[期刊论文]-计算机工程与设计 2010(1)

6. 张丽敏 垂直搜索引擎的主题爬虫策略[期刊论文]-电脑知识与技术 2010(15)

7. 曾伟辉. 李淼. 曾伟辉 深层网络爬虫研究综述[期刊论文]-计算机系统应用 2008(5)

8. 荣光. 张化祥 一种Deep Web爬虫的设计与实现[期刊论文]-计算机与现代化 2009(3)

9. 鞠彦辉. 许燕 Deep Web信息资源开发策略研究[期刊论文]-现代情报 2008(1)

10. 黄聪会. 张水平. 胡洋 主题Deep Web爬虫框架研究[期刊论文]-计算机工程与设计 2010(5)

11. 孙彬. 王东. 李娟 基于XQuery的Deep Web搜索系统的设计与实现[期刊论文]-科学技术与工程 2007(16)

12. 刘汉兴. 刘财兴 主题爬虫的搜索策略研究[期刊论文]-计算机工程与设计 2008(12)

13. 张鑫. 陈梅. 王翰虎. 王嫣然 基于视觉特征和领域本体的Web信息抽取[期刊论文]-计算机技术与发展 2011(2)

14. 陈方. 谭爱平. 成亚玲. 文益民 主题爬虫技术研究综述[期刊论文]-湖南工业职业技术学院学报 2008(5)

15. 孙立伟. 何国辉. 吴礼发 网络爬虫技术的研究[期刊论文]-电脑知识与技术 2010(15)