深层网络爬虫研究综述①

Survey on the Research of Deep Web Crawler

曾伟辉　李　淼　（中国科学院合肥智能机械研究所 安徽合肥 230031）
曾伟辉　（中国科学技术大学 自动化系 安徽合肥 230027）

摘　要： 随着 Internet 的迅速发展，网络资源越来越丰富，人们如何从网络上抽取信息也变得至关重要，尤其是占网络资源 80% 的 Deep Web 信息检索更是人们应该倍加关注的难点问题。为了更好的研究 Deep Web 爬虫技术，本文对有关 Deep Web 爬虫的内容进行了全面、详细地介绍。首先对 Deep Web 爬虫的定义及研究目标进行了阐述，接着介绍了近年来国内外关于 Deep Web 爬虫的研究进展，并对其加以分析。在此基础上展望了 Deep Web 爬虫的研究趋势，为下一步的研究奠定了基础。

关键词： 深层网络 网络爬虫 信息检索 URL 动态网页

1　引言

随着人们对信息重要性认识的加深，信息获取方式也越来越多，作为各种信息的载体，网络蕴含着大量的资源，如何快捷的从网络上获取所需信息成为人们亟待解决的问题。各种搜索引擎应运而生，例如传统的通用搜索引擎 AltaVista、yahoo！、google 等。而这些搜索引擎存在着一定的局限性，尤其是它们无法搜索到深层网络（Deep Web）的信息。据 BrightPlanet 公司技术白皮书[1]，Deep Web 里包含的可访问信息容量是一般的 Surface Web 的 400 – 500 倍。可见，研究 Deep Web 爬虫对于提高搜索覆盖率和准确率有着非常重要的意义。

2　深层网络爬虫的定义和研究目标

网络爬虫，又称网络机器人。英文名有 Spider、Crawler、Bots、Robot、Wanderer 等。它是一个自动提取网页的程序，是搜索引擎的重要组成部分[2]。

互联网网页按存在方式可分为“表层网”（Surface Web）和“深层网”（Deep Web，也有称 Invisible Web，HiddenWeb）[3]。Surface Web 指传统网页搜索引擎可以索引的页面，以超链接可以到达的静态网页为主构成的 Web 页面。Deep Web 是指那些存储在网络数据库中，不能通过超链接访问而通过动态网页技术访问的资源集合。它最初由 Dr. Jill Ellsworth 于 1994 年提出，定义为那些由普通搜索引擎难以发现其信息内容的 Web 页面[4]。

Christ Sherman 等人[1]对 Deep Web 定义为：虽然通过互联网可获取，但是普通搜索引擎受技术限制不能或不作索引的那些文本页、文件或其它高质量的、权威的信息。

文献[5]对 Deep Web 定义为：那些大部分内容是不能通过静态链接获取的，特别是大部分隐藏在搜索表单后的，只有用户键入一系列关键词才可获得的页面。

广义上来说，Deep Web 包含四个方面[6]：(1) 通过填写表单形成对后台再现数据库查询得到的动态页面。(2) 由于缺乏被指向的超链接而没有被索引到的页面。(3) 需要注册或其他限制访问的页面。(4) 可

① 基金项目 中科院知识创新工程重要方向项目（KGCX2 – SW – 511）

访问的非网页文件。

因为担心爬虫会陷入海量动态网页库而浪费网络带宽和存储资源,而且目前的技术还无法发现潜藏在网络数据库中的信息,所以传统搜索引擎,比如 AltaVista、yahoo！、google 等,一般只索引 Surface Web 中静态网页、文件等资源,却不索引或很少索引 Deep Web 中的资源。

2000 年 Bright Planet 公司针对 Deep Web 作了一个详细的调查,发布了调查白皮书,提到几点调查发现[1]:(1) Deep Web 中可访问信息容量是 WWW 的 400－500 倍;(2) Deep Web 包含 7500TB 的信息,而 Surface Web 中只有 19TB;(3) Deep Web 包含近 5500 亿独立文档,而 Surface Web 只包含 10 亿;(4) 现有 Deep Web 站点已超过 100,000 个;(5) 最大型 Deep Web 站点的 60% 所包含资源数量约有 750GB,已是 Surface Web 的 50 倍;(6) Deep Web 站点月访问量是一般站点的 150%,并且经常被链接;(7) Deep web 是互联网上最大、发展最快的新型信息资源;(8) Deep Web 站点比一般站点涉及范围较小,内容更为精深;(9) Deep Web 包含的有效高质内容总量至少是 Surface Web 的 1000—2000 倍;(10) Deep Web 的信息内容与所有的信息需求、市场和领域高度相关;(11) 超过一半的 Deep Web 内容都保存在专业领域数据库中;(12) 95% 的 Deep Web 信息都可被免费访问。

在 2004 年,UIUC 大学又对 Deep Web 作了一次估算[7],推测出整个 Web 上有 307000 个网络数据库站点,450000 个网络数据库,比 Bright Planet 估计的 50000 个又翻了许多倍。

通过上述研究可见,研究 Deep Web 爬虫是进一步提高互联网信息获取质量和数量的有效途径。但是,通用网络爬虫在处理 Deep Web 时通常会遇到如下问题[8]:

(1) 不具备处理浏览器端脚本代码中所有可能产生新的浏览器导航信息的机制。

(2) 不具备处理会话持久的相关机制,例如 cookie。

(3) 许多站点使用了复杂的重定向技术。

(4) 不能处理其他客户端技术,如 java applet,flash 等。

(5) 一些复杂的 HTML 代码,例如 frames 嵌套,https 技术等。

3 深层网络爬虫研究进展

现有的 Deep Web 爬虫技术大部分是基于表单填写,按表单填写方法可分为两类:1) 基于领域知识的表单填写。该方法一般都有一个本体库,通过语义分析来选取合适的关键词组合填写表单;2) 基于网页结构分析的表单填写。此方法一般无领域知识或者仅有有限的领域知识,将网页表单构建成 DOM 树,在 DOM 树中提取表单各字段值。此外,还有一些爬虫技术能够处理 javascript 语言。

3.1 基于领域知识的爬虫技术

Raghavan S 等人[9] 提出的 HIWE 系统中,Crawler 管理器负责管理搜集过程。它对下载的 Web 页面进行分析,包含表单的页面被送到表单处理器处理。表单处理器先从页面中抽取出表单,再从预先准备好的数据集中选择数据自动完成填写,然后将合成的 URL 提交给 Crawler 管理器以下载相应的结果页面。该方法要求用户事先准备相应的表单数据集,每个表单项只跟一个文本相关联,不能站在全局的观点上来处理表单项,且不能处理 javascript 脚本。

Yiyao Lu[10] 等人提出一种获取 Form 表单信息的多注解方法。该方法首先将数据单元按语义分配到各个组中,接着对每组从多方面注解,集合各种注解结果来预测一个最终的注解标签。但当在相同的 SRR 中的一个属性有多重数据单元时(例如一本书有多个作者),该方法将会出错。

Zhen Zhang 等人在文章[11] 中提出一种轻量级的基于领域知识的自动表单填写框架。其核心是一种基于类型的搜索驱动方法,此方法能将查询重定向到一组相关资源的集合。

严亚兰[12] 提出了一种面向动态网页爬行的 Crawler 架构。它包括 crawl 模块、表单分析模块、表单处理模块、结果分析模块、语义词表管理模块、URLs 集和语义词表,并与搜索引擎中的索引数据集或网页数据集发生信息交互。crawl 模块控制并执行所有的爬行过程。它先从一个种子 URLs 集开始爬行,对爬回的静态

网页,进行以下几方面的处理:从网页中抽取所有的由链接指向的 URLs,并将这些 URLs 存入到 URLs 集中;将爬回的网页保存在网页数据集中(如果存在),或者保存在缓存中足够长的时间,使索引模块完成索引任务并将索引数据保存在索引数据集中,并保证表单分析模块完成对网页表单的分析。

郑冬冬等人[5]提出的爬虫,首先针对站点接口产生一个查询;然后下载结果索引列表页面;最后根据结果的索引下载具体页面。其中查询关键字选择策略采用三种方式:随机选择、根据词频选择和适应性策略方式。查询选择算法主要是通过 Efficiency(qi) = P(qi)/Cost(qi)来计算查询 qi 单位代价下获取新的文档页面数,以此获得效能值最大的关键词。

Alvarez 等人[13]提出的 Deep Bot 是基于一系列的领域知识,每种领域定义描述了一个数据采集任务。它采用了基于视觉距离和角度的方法计算表单各项的最佳项值,距离超出阈值的文本舍弃,距离相等的文本比较该文本与表单项的角度,表单项与文本之间的文本相似度通过文献[14]中的 TFIDF 和 Jaro – Winkler edit – distance 算法来计算。并且通过各表单项的相似度计算表单与主题的相似度,来识别与主题相关的表单,模拟填写表单,基于对象和客户端脚本的事件触发机制提交执行表单来得到 hidden web 中隐藏的信息。该方法不适于快速搜索和加密的表单。

3.2 基于网页结构分析的爬虫技术

Bergholz 等人[15]针对全文本搜索表单的 Deep Web 进行处理,这种网页的搜索表单只有一个输入关键字的表单项。Alexandros Ntoulas 等人在文献[16]中可以针对前面的搜索结果自动产生新的搜索关键词,并对它们进行优先级排序以获得隐藏在表单后的信息。该方法优点是可以使用最少的提交次数得到用户想要的数据,但是不能处理含多形式的表单。

Ali I. El – Desouky 等人[17]提出一种 LEHW 方法,该方法通过一个解析器(将 HTML 网页表示成一个 DOM 树形式)来区分(S – A)和(M – A),索引表单通过两种不同的数据结构,一种是针对 S – A 表单,另一种是针对 M – A 表单,然后分别通过页面标签判断进行处理。此方法不仅可以对 multi – attribute(M – A)型表单进行处理,而且对现有的 single – attribute (S – A)型表单处理技术改进。但不足之处在于 LEHW 方法在提取标签方面精确度没有 HIWE[9]方法高。

陈珂等人[18]提出一种两阶段采样策略,确定是否充分获取了后台数据库数据。首先用默认值来提交表单,然后对表单元素值组合进行采样以确定该提交是否返回了后台数据库的所有数据,若是,则结束提交过程;否则,在爬虫所具有资源限制范围内穷尽所有可能值的组合。该方法不足之处是只能获取一部分 Deep Web 页面,且无法处理文本域元素。

Luciano Barbosa 等人[19]提出了一种自适应抓取策略,以有效地找到切入点隐藏 web 资源。由于隐藏网页的来源是稀疏分布的,作者通过网页内容将爬行定位到某一主题,优先考虑主题相关链接,并跟踪那些可能不会导致立即受惠的链接,并提出了一个新框架,以自动学习模式来调整爬行方向,大大减少了手工设置和调整。由此,我们可以在设计爬虫时适当考虑那些非主题链接,在选取链接时,自动学习模式是值得借鉴的。

孙彬等人在文章[20]中提出一种基于 XQuery 的搜索系统。该系统模拟表单和特殊页面标记切换,把页面关键字切换信息描述为三元组单元,按照一定规则先进行盲搜索,排除无效表单,然后将 Web 文档构造成 DOM 树,利用 XQuery 完成文字属性映射到表单字段的识别过程。此方法在处理常规结构的站点时其爬行覆盖率达到了 71% 以上。但是对于现在大量出现的非常规结构的站点,没有很好的处理。

宋晖等人[21]提出了一套自动查询 Hidden Web 信息的系统 HWIR。用户可以输入索引主题及相关文本来检索包含了 Hidden Web 信息的 Web 页面。HWIR 利用 Crawler 来收集网页,使用对象抽取技术从网页中分析 Hidden Web 中数据库的人口 Form 表单,然后自动建立 Hidden Web 信息的索引,用户可通过结构化查询获取所需的 Hidden Web 信息的网页。文中采用的 TTOE 技术是一种基于标记树结构的表单抽取算法。它首先将 Web 页面表示成树型结构,然后再以此树为基础进行表单对象的抽取。该方法抽取对象主要针对 HTML、XML 中的文本信息,对用 ASP、JSP 的代码没有分析,而很多的 Hidden Web 信息入口就隐藏在这些代

码中。

3.3 基于脚本语言分析的爬虫技术

目前基于脚本语言的爬虫技术,通用的方法[22]是用脚本分析引擎来模拟浏览器动作,执行脚本代码。开放源码的 JavaScript 引擎 SpiderMonkey 提供了一个最基本的且易于扩展的 JavaScript 分析器。通过包装 SpiderMonkey,使其接口能接收从页面提取的 JavaScript 代码,返回执行 JavaScript 后得到的所有 URL,从而完成爬虫任务。

Alberto Pan 等人提出另一种解决客户端隐藏网络的解决方案[6]。该方案包含三个步骤:1)将网页文档当作路由表来处理会话持久问题。2)使用标准浏览器 API 自动构建迷你浏览器替代 http 客户端处理脚本执行代码,页面重定向。3)通过一种自底向上的递归算法来处理弹出菜单,以及其他动态页面元素。

上面的两种解决方案为我们着手脚本代码的研究指明了方向,我们可以在此基础上进一步的改进与完善。

4 深层网络爬虫的研究趋势

本文从网络信息生产的趋势看,越是价值高、规模大的信息往往越深藏在深层网络中,而现在大部分的网络爬虫都无法对深层网络中的 Flash 和 Script 等动态网页和数据库进行采集。当前对于 Deep Web 爬虫技术的研究大多只是针对表单的,少数针对 javascript 脚本代码的 Deep Web 处理[22]。深层网络爬虫的研究将趋向于以下几个方面:

一 关于 AJAX 技术的深层网络爬虫研究。

AJAX 技术现已被广泛使用在网页中。Google 的 Orkut,Gmail,以及最近的 beta 版的 Google Groups、Google Suggest 和 GoogleMaps,都应用了 AJAX 这项技术。Flickr、Amazon 的 A9.com 搜索引擎也采用了类似的技术。Microsoft 已经推出了 Atlas 的 β 版,它在 ASP.NET 中实现了 Ajax。BEA Systems 公司正在把 Ajax 功能构建到它的门户产品中并把 Ajax API 加入运行时工具。Sun Microsystems 公司计划把 Ajax 加入 Java Server Faces。企业服务总线供应商 CapeClear Software 公司则计划把 Ajax 工具加入以 SOA 为中心的产品中。Ajax 的广泛研究和应用,使得以 ajax 为基础的新一代 javascript 网络站点信息抽取问题的研究显得越来越重要。

通过我们对国内外研究进展的探讨与分析,目前很少有人针对此类网页进行爬虫技术的研究[23]。这方面的研究也将成为 Deep Web 爬虫需要处理的技术难点之一。

二 多媒体网络爬虫研究

随着超媒体技术和宽带网技术的发展,开发可查寻图片、声音、图像和电影的搜索引擎是一个新的方向[24]。因特网上图形、图像、视频 音频、动画等多媒体信息正日渐丰富。同时,用户对其检索的要求也在不断增长。伴随着搜索引擎的发展,各种基于网络的多媒体爬虫技术研究将会成为爬虫研究的新方向。

三 对等网络 p2p(Peer – to – peer)

对等网络在加强网络上人的交流、文件交换、分布计算等方面大有前途。长期以来,人们习惯的互联网是以服务器为中心,人们向服务器发送请求,然后浏览服务器回应的信息。而 p2p 所包含的技术是使联网电脑能够进行数据交换。但数据是存储在每台电脑里,而不是存储在既昂贵又容易受到攻击的服务器里。网络成员可在网络数据库里自由搜索、更新、回答和传送数据。很多人都共享了他们认为最有价值的东西,这将使互联网上信息的价值得到极大的提升。

5 结束语

随着 web2.0 网站的大量涌现,通用网络爬虫技术日臻成熟,面向深层网络的爬虫技术已经开始成为搜索引擎发展的主要趋势之一。本文在给出深层网络的定义,以及通用网络爬虫在深层网络中遇到的困难与挑战后,对现有的各类面向深层网络的爬虫技术进行了分析,希望通过本文能够对这一领域的研究有一个比较清晰的概括与总结。

总之,针对深层网络的爬取技术仍然处于探索阶段,距离实际应用还比较远,仍有大量的问题需要我们去研究。

参考文献

1 MICHAEL K. BERGMAN, The Deep Web: Surfacing

Hidden Value. http://www.completeplanet.com/Tutorials/DeepWeb/[EB/OL], 2000.

2 刘金红,陆余良. 主题网络爬虫研究综述. 计算机应用研究,2007,24(10):26-29、47.

3 李涛,陈鹏,李哲. 深度 Web 资源探测系统的研究与实现. 微计算机信息,2007,23(11-3):185-187.

4 郑冬冬,赵朋朋,崔志明. DeepWeb 爬虫研究与设计. 清华大学学报(自然科学版),2005,45(S1):1896-1902.

5 郑冬冬,崔志明. Deep Web 爬虫爬行策略研究. 计算机工程与设计,2006,27(17):3154-3158.

6 Manuel Alvarez, Alberto Pan, Juan Raposo, Angel Vina. Client-Side Deep Web Data Extraction extended paper, http://www.tic.udc.es/~mad/publications/csdeep web_extended.pdf, 2002.

7 Chang K C C, He B, Li C, et al. Structured databases on the web: Observations and implications[C] SIGMOD Record, 2004,33(3).

8 Manuel Alvarez, Alberto Pan, Juan Raposo, Angel Vina. Client-Side Deep Web Data Extraction. Proceedings of the IEEE International Conference on E-Commerce Technology for Dynamic E-Business.

9 Raghavan S, Garcia-Molina, H. Crawling the Hidden Web. Report, 2000(36), http://dbpubs.stanford.edu/pub/2000-36).

10 Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng. Annotating Structured Data of the Deep Web, IEEE, 2007:376-385.

11 Zhen Zhang, Bin He, Kevin Chen-Chuan Chang. Light-weight Domain-based Form Assistant: Querying Web Databases On the Fly. In Proceedings of the 31st Very Large Data Bases Conference, 2005: 97-108.

12 严亚兰. 面向动态网页爬行的 Crawler 架构. 图书情报知识, 2003(4):51-53.

13 Manuel Alvarez, Juan Raposo, Alberto Pan, Fdel Cacheda, Fernando Bellas, Victor Carneiro. DeepBot: A Focused Crawler for Accessing Hidden Web Content. ACM, 2007:18-25.

14 Cohen, W, Ravikumar., P., Fienberg, S. A Comparison of String Distance Metrics for Name-Matching Tasks. In Proceedings of IJCAI-03 Workshop. 2003:73-78.

15 Bergholz A, Chidlovskii, B. Crawling for Domain-Specific Hidden Web Resources. Conference on Web Information Systems Engineering. 2003:125-133.

16 Ntoulas, A., Zerfos, et al. Downloading Textual Hidden Web Content Through Keyword Queries. Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries. 2005:100-109.

17 Ali I. El-Desouky, Hesham A. Ali, Sally M. El-Ghamrawy. An Automatic Label Extraction Technique fo-r Domain-SpecifIc Hidden Web Crawling (LEHW), IEEE2006:454-459.

18 陈珂,陈小英,徐科. Hidden Web 信息获取. 计算机时代,2007(5):54-56.

19 Luciano Barbosa, Juliana Freire. An Adaptive Crawler for Locating Hidden-Web Entry Points. Banff, Alberta, Canada. WWW2007, 5:441-450.

20 孙彬,王东,李娟. 基于 XQuery 的 Deep Web 搜索系统的设计与实现. 科学技术与工程,2007,7(16):4080-4084.

21 宋晖,张岭,叶允明,马范援. 基于标记树对象抽取技术的 Hidden Web 获取研究. 计算机工程与应用,2002,(23):9-12,24.

22 王映,于满泉,李盛韬,王斌,余智华. JavaScript 引擎在动态网页采集技术中的应用. 计算机应用,2004,24(2):33-36.

23 罗兵. 支持 AJAX 的互联网搜索引擎爬虫设计与实现[硕士学位论文]. 杭州,浙江大学,2007.

24 彭建荣,罗永会. 搜索引擎的基本原理及发展趋势. 电脑知识与技术,2006(2):84-85.