Web数据抽取技术研究进展

张成洪　古晓洪　白延红
（复旦大学管理学院　上海200433）

摘　要　由于Web上存在着大量有用而复杂的信息，近年来学术界和企业界开发了许多从Web中抽取数据的方法和工具。本文总结了Web数据抽取技术的研究进展和从Web中抽取数据的主要原理、过程、方法和抽取规则，并讨论了未来的研究方向。

关键词　Web数据抽取，Web查询，包装器，抽取规则

The Progress of Web Data Extraction Technology

ZHANG Cheng-Hong[1,2]　GU Xiao-Hong[1]　BAI Yan-Hong[1]
(Department of Information Management & Information System, Fudan University, Shanghai 200433)[1]
(Informatization Office, Fudan University, Shanghai 200433)[2]

Abstract　In order to extract data from the Web, a great number of approaches and tools have been developed by academic community and commercial companies. This article gives an overview of the progress of Web data extraction technology and outlines the main principles, processes, approaches and extraction rules to extract data from the Web. It also points out the direction for the research in the future.

Keywords　Web data extraction, Web query, Wrapper, Extraction rules

1　引言

Web上存在着丰富的信息，企业的很多信息也主要以HTML格式存在于Web上。最初Web数据获取方法的研究主要集中在两个方面：搜索引擎和Web查询（或半结构化数据查询）。搜索引擎主要是通过关键字进行查询，但效率较低，存在诸多问题[1]。Web查询研究的基本思想则是利用数据库或数据仓库的概念，把Web看成是一个由非结构化和半结构化文档构成的大型分布式数据库系统，通过构建基于其上的超文本关系视图，实现Web结构和内容查询。这类研究包括两部分：构建针对Web的数据模型（例如OEM[2,20]、UnQL[3]、Strudel[4]、NSL[5]等）和提供数据库形式的针对Web的查询语言或系统（例如W3QS[6]、WebSQL[7]、WebLog[8]、UnQL[3]、Lorel[9]、WebOQL[10]、Strudel[4]、Florid[11]、WebML[12]、HTML-QL[13]等）。这种方法也存在很大的不足，主要是其查询结果粒度较粗且不够准确。

目前出现了一些全新的基于Web的应用[14]，例如监控股票市场的及时行情、比较各个站点的商品价格、跟踪合作伙伴或竞争对手的动态、集成企业内部和外部的各种信息包括位于Web上的信息等。此外，人们希望利用智能代理技术或专门的应用程序以自动完成某些工作。因此，提供的数据必须是可信赖（reliable）且机器可读的（machine readable），而仅凭搜索引擎和Web查询技术显然无法满足这样的要求。

Web数据抽取技术已成为当今研究的一个热点。Web数据抽取及其数据管理方法，可以解决上面提到的一些问题。利用数据抽取技术，通过对特定网页的结构和数据项特征进行分析，可以将网页中感兴趣的信息准确地抽取出来，并保存到数据库系统或其他格式的文件中（例如XML、OEM[2]等），以供SQL语言、半结构化查询语言或XML查询语言[15,16]查询，或者供其他应用程序使用。

目前Web数据抽取技术广泛采用的是称为网页包装（wrapping）的方法，主要研究集中在抽取方法的研究[17]和wrapper的自动或半自动化生成工具的研究上。抽取方法主要包括直接解析[18,19]、HTML结构分析[21,22]和数据建模[23,24]。关于wrapper生成工具，最初的wrapper程序主要是采用手工编写[25~27,18]，这种方法存在很多缺点，主要是开发和维护都很困难。鉴于此，出现了很多wrapper的自动或半自动化生成工具。其中，Araneus[28]、BYU[29]、DEByE[30]、Jedi[31]、NoDoSE[32]、Road Runner[33]、WebL[34]、Xwrap[35]等是非商业工具，主要由大学的研究团体开发。而商业工具则包括RoboSuite[36]、Visual Web Task[37]、X-Fetch Wrapper[38]、W4F[39]、Info Scanner[40]等。文[41,42]从不同角度对这些工具进行了分类。

本文主要对Web数据抽取的基本原理和目前存在的主要抽取方法进行归纳总结。第2部分给出了一个简化的Web数据抽取模型，简要介绍了wrapper的构成及Web数据抽取的基本步骤；第3部分介绍了Web数据抽取的主要方法；第4部分讨论了抽取规则；最后展望了Web数据抽取的发展趋势。

2　Web数据抽取的一般模型和基本步骤

为了抽取特定网站的数据，主要方法是要构建一个基于该信息源的通用的数据抽取模型。虽然数据抽取模型的定义、结构和使用各异，但都是针对特定的网页或网站进行处理，广

张成洪　副教授，主要从事信息系统规划、数据仓库、企业门户、Web数据抽取等方面的研究。古晓洪　硕士研究生，主要从事语义网、Web数据抽取等方面的研究。白延红　硕士研究生，主要从事金融工程方面的研究。

义上都可以称之为网页包装(wrapper)。但通常说来,一个wrapper主要由抽取规则(extraction rules)和抽取器(extractor)两部分构成。抽取规则主要描述网页结构、数据项位置、抽取步骤、转换规则、输出方式等。而抽取器是一个可执行程序或其他应用程序(例如XML应用程序),用来执行抽取规则,产生结果数据。有的wrapper系统还包括超链分析、数据校验、数据映射等[21]。此外,一个完整的Web数据抽取系统还包括数据集成(Data Integration)功能[21]。

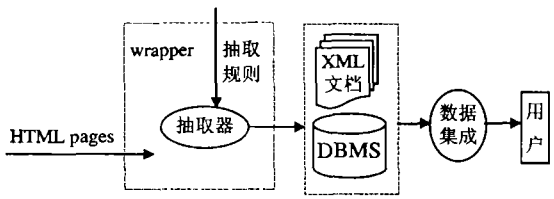

图1 Web数据抽取模型

一个典型的Web数据抽取过程主要包括下列五个步骤:网页获取、数据抽取、数据校验、数据存储和数据集成。

1. 网页获取 也就是获取目标网页。用户可以指定一个网址或若干网址,或者制定导航规则,由抽取程序按规则自动爬行并获取网页。根据数据抽取的特点,网页可以分两类:一类是数据页面,包含需要抽取的数据;一类是导航页面(Navigational Pages),包含指向数据页面的若干链接[21]。导航页面一个典型例子是搜索引擎的返回结果。在一个搜索结果页面总是包含若干相似结构的链接信息,分别指向不同的目标网页。有的系统不仅能跟踪静态超链接,而且还能访问"deep Web"[43],发现隐藏在Web表单和JavaScript代码中的动态链接[21,44]。

2. 数据抽取 获取目标网页后,下一步骤就是数据抽取。数据抽取是一个wrapper完成的核心功能,主要任务是根据抽取规则解析HTML文件,抽取需要的数据项,或者先对原始HTML文件进行预处理(例如首先转换为XHTML格式文件),再对预处理后的文件进行处理。

3. 数据校验 为了保证数据抽取的质量,尽量减少"脏数据",取得的数据一般还要经过数据校验步骤才会输出。数据校验存在两个层次。首先是确保抽取数据本身的正确性和完整性。对于一些明显的错误,可以通过XML模式文件、规则表达式等手段加以限制,通过应用领域知识和机器学习等技术加以纠正或将丢失的数据补充完整。其次,当从多个Web站点抽取并集成(aggregating)数据时,不同的Web站点可能遵循不同的命名规范和采取不同的计量单位,需要将数据映射到一个标准格式以改善抽取数据质量,这可以通过在代码中包含条件声明、规则表达式和领域知识,或者结合数据库技术来完成。

4. 数据集成 Web数据抽取的最后一个任务是从多个相关的Web页面中集成数据。对单个站点而言,一些Web站点采用HTML框架进行页面布局,或者为了不在一个页面中充斥过多的数据而将一些相关的数据部署到多个"兄弟"页面(sibling pages)上,从而将一个逻辑数据单元分割成多个分开的HTML文档。对多个站点而言,用户为了完成某些功能,例如产品价格比较,也需要将不同网站的内容进行综合。

3 Web数据抽取方法的比较

Web数据抽取方法实际上取决于wrapper的构造机制。本文主要根据wrapper中数据抽取器的实际运行机制,将目前研究文献中出现的主要Web数据抽取方法归为三类。

3.1 直接解析HTML文档的方法

该方法利用Perl、Java、YACC、Phyon等程序语言或其他自行设计的程序语言,编写可执行程序直接对HTML网页进行分析和处理。这种方法主要利用规则表达式对内容进行模式匹配,不涉及HTML文档的层次结构。这种方法有很多众所周知的缺点,主要的是程序的健壮性和可维护性较差。因为抽取规则固化在程序中,一旦网页内容和结构发生变化,就必须对wrapper进行重新设计。

后来出现的一些Web数据抽取方法中引入了规则文件的概念。抽取逻辑从程序中被分离出来放入规则文件中,一旦结构发生变动,或者需要抽取同类网页数据,只需改写规则文件。这在很大程度上弥补了上面谈到的缺陷。规则文件以各种形式存在着,例如描述文件(specification files)[18,19]、XSLT文件[21]、DEL脚本文件[38]等。

3.2 分析HTML层次结构的方法

这种方法实际上主要是利用XML技术,因此也可称作基于XML的方法。随着XML技术的出现,XML已成为Web上重要的数据表示和交换标准。因此,Web数据抽取就不能不考虑到Web上将会出现大量的XML文档和利用强大且日益成熟的XML技术。基于XML的Web数据抽取也已经成为一种趋势。

该方法首先将HTML文档根据DOM转换为一棵具有层次结构的HTML树。基本做法是将原始的HTML文档通过一个过滤器(filter),该过滤器检查并修改HTML文档的语法结构,从而形成一篇良构(well-formed)的HTML文档,即XHTML。由于XHTML是基于XML的,因此下一步就可以利用XML工具来处理这些HTML文档。实现这一转换步骤的工具有很多,包括Tidy[45]、W4F[5]等。

下一步骤就是处理XHTML文档。如果抽取的结果是XML文档,实际上可以把将目标XHTML转换为结果XML文档的这一过程视为一个XML转换问题。目前存在几种不同的转换方式。

ANDES[21]选择的数据转换机制是XSLT(Extensible Stylesheet Language Transformation)。XSLT及其扩展机制提供了强大XML路径表达式(XML path expressions,XPath)和规则表达式(regular expressions)能力。ANDES中,一组XSLT文件扮演着规则文件的角色,定义了抽取内容、步骤和校验规则。XHTML文档依次通过一组XSLT文件,最后输出一个XML文件,其结构和内容由最后一个XSLT文件决定。

相对于XSL,DEL(Del Extraction Language)语言是另一种值得注意的工具。DEL语言是Republica公司[46]为其产品X-Fetch Wrapper[38]开发的语言,已经被W3C添加到其note list中[47]。基于XML的DEL语言的优点是能够快速方便地完成任何格式到XML转换过程的描述。

3.3 基于概念建模的方法(Conceptual-Model-Based Approach)

该方法主要基于ontology概念。BYU Data Extraction Group[29]对此进行了大量研究。该方法先用ontology建立数据模型,再把可能抽取的数据项映射到ontology中的元素上,用户选择Ontology中的元素以决定抽取的对象。Ontology的引入既保证了结构的一致性,又保证了数据的一

致性，使不同来源的数据都能以统一的视图呈现，方便了信息的继承和交换。

4 Web 数据抽取中规则的表示

Wrapper 的另一主要组成部分是抽取规则。前面提到，抽取规则包含在规则文件中，规则文件可以以各种形式存在。抽取规则中最主要的规则是关于数据项的定位问题，即如何快速、准确地定位到网页中相关的数据。数据项的定位方式在很大程度上影响到数据抽取的质量。下面对一些主要数据项定位方法进行比较。

1. 模式匹配法：模式通常用规则表达式来表达。利用模式和规则表达式，通过分析数据项的前后边界来定位数据项。模式通常发挥着过滤器的作用，不适合模式的部分被自动抛弃。一般具有强大字符串处理能力的高级语言，例如 PERL，均可以采用这种方法进行数据抽取。

2. 绝对路径法：绝对路径从 HTML 树的顶部节点(〈HTML〉标签)开始。象 W4F，其 HTML Extraction Language 需要使用绝对 HTML 路径。然而，当目标 HTML 页改变后，绝对路径法也就失效了。HTML 设计最常改变的是数据项在页面中的位置。当向页面添加了新内容或者移动了原先的内容，HTML 标签的绝对位置也发生了变化。因此，建立起独立于数据项绝对路径的位置表示就很重要。

3. “锚点”(anchor)法：在页面中确定一锚点，作为路径表达式的起点。锚点的选取应该基于数据项内容而不是其路径。例如，在给出一本书的价格时，其前面通常有“价格”这个词。通过首先找到“价格”这个词，就为价格数据项建立了一个独立于其绝对路径的锚点。一个页面可以有一个总的锚点，各个数据项的位置从这个锚点开始描述。也可以为每一个数据项分别建立自己的锚点。

4. 有些系统(如 Road Runner[33])：没有独立的规则文件，因此也并不在规则文件中预先确定数据项位置。其抽取原理是在抽取过程中比较两个或多个预先给定的属于同一类的样本页面的结构，从而自动生成页面中所包含数据的模式(schema)。

现在有越来越多的工具，利用机器学习、自然语言处理等技术，可以自动确定抽取规则和生成规则文件。这些工具在引言中已有叙述。

5 Web 数据抽取的研究趋势

目前 Web 数据抽取基本上基于这样一些假设：1)信息有规律地更新，但数据结构保持不变；2)很多 Web 站点实际上采用一组 HTML 模板和后端数据库进行驱动。应该说，这些假设基本上是成立的，因此 Web 数据抽取技术也得到了广泛的应用。然而，HTML 页面内容和结构确实会发生改变，一旦出现这种情况，就必须重新生成 wrapper 或规则文件。为了有效解决这个问题，Web 数据抽取存在下列发展趋势。

Web 数据抽取的一个发展趋势在抽取过程中融入更多的自然语言处理、机器学习等技术和利用图形化工具，以提高抽取的自动化程度和数据抽取质量，简化抽取过程。

Web 数据抽取的另一个发展趋势是是利用语义网(Semantic Web)[48]技术。通过在网页本身嵌入一些语义标记，抽取程序就能根据这些语义标记识别并抽取相关内容。WAP 站点为不同型号的手机提供不同的内容，原理就在于在网页中嵌入不同的标记，可以看作是这种方法的一种应用。由于需要操纵服务器端网页本身，相对其他方法需要在客户端处理而言，这可以看作一种服务器端抽取模式，也无疑是最理想的一种方式。这种方法的缺点是目前采用语义网技术的网页数量还很少。

参 考 文 献

1 Brin S, Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Proc. of the 7th WWW Conf. 1998. 107～117
2 Papakonstantinou Y, Garcia-Molina H, Widom J. Object Exchange Across Heterogeneous Information Sources. In: Proc. of the Intl. Conf. on Data Engineering, IEEE Computer Society, 1995. 251～260
3 Buneman P, Davidson S, Hillebrand G, Suciu D. A query language and optimization techniques for unstructured data. In: Proc. of SIGMOD'96, June 1996
4 Fernandez M, Florescu D, Lery A, Suciu D. A Query language for a Web-Site Management System. SIGMOD Record, 1997. 4～11
5 Sahuguet A, Azavant F. Building Light-Weight Wrappers for Legacy Web Data-Sources Using W4F. In: Proc. Intl. Conf. on Ve-ry Large Data Bases (VLDB), Edinburgh, Scotland, Sep. 1999
6 Shmueli O, Konopnicki D. W3QS: A Query System for the World-Wide Web. In: proc. of the Intl. Conf. on Very Large Data Bases, 1995. 54～65
7 Mendelzon A, Mihaila G, Milo T. Querying the World Wide Web. In: proc. of the First Intl. Conf. on Parallel and Distributed Information System, 1996. 80～91
8 Lakshmanan L V S, Sadri F, Subramanian I N. A declarative language for Querying and Restructuring the Web. In: Proc. of the 6th Intl. Workshop on Research Issues in Data Engineering, 1996
9 Abiteboul B, Quass D, McHugh J, Widom J, Wiener J L. The Lorel Query Language for Semistructured Data. Intl. Journal of Digital Libraries, 1997, 1(1): 68～88
10 Aro-cena G, Mendelzon A. WebOQL: Restructuring Documents, Databases and Webs. In: Proc. of the Intl. Conf. on Data Engineering, IEEE Computer Society, 1998. 24～33
11 Himmero-der R, Lausen G, Ludascher B, Schlepphorst C. On a declarative semantics for web queries. In: Proc. of Intl. Conf. on Deductive and Object-Oriented Databases, 1997. 386～398
12 Zaiane O, Han J, Li Z, Chee S, Chiang J. WebML: Querying the World Wide Web for Resources and Knowledge. In: Workshop on Web Information and Data Management, (WIDM'98), 1998
13 Liu M, Ling T M. A Conceptual Model and Rule-based Query Language for HTML. World Wide Web, 2001, 4(1): 49～77. http://www.scs.carleton.ca/～mengchi/papers/WWWJ01.ps
14 Brin S, Motwani R, Page L, Winograd T. What can you do with a Web in your pocket? Data Engineering Bulletin, 1998, 21(2): 37～47
15 Chamberlin D D, Robie J, Florescu D. Quilt: An XML Query language for Heterogeneous Data Sources. In: Proc. of the Third Intl. Workshop on the Web and Databases, Dallas, Texas, U. S. A., May 2000. 53～62
16 Deutsch A, Fernandez M F, Florescu D. Levy A, Suciu D. XML-QL: A Query Language for XML. Submission to W3C. http://www.w3.org/NOTE-xml-ql-19990819, 1998
17 Eikvil L. Information Extraction from World Wide Web -- A Survey: [Report No. 945, ISBN 82-539-0429-0]. July 1999
18 Chawathe S, et al. The TSIMMIS Project: Integration of Heterogeneous Information Sources. In: Proc. IPSJ Conf. 1994
19 张成洪，肖军建，张诚. Web 内容抽取及其数据管理方法. 复旦学报，2001(2)
20 黄豫清，戚广智，张福炎. 构造 Web 文档中半结构化信息技术. 计算机辅助设计与图形学学报，2000，12(3)
21 Myllymaki, Jussi. Effective Web Data Extraction with Standard XML Technologies. International Journal of Computer and Telecommunications Networking. In: 10th Intl. World Wide Web Conf. Hong Kong, May 2001
22 Sahuguet A, Azavant F. Looking at the Web through XML glasses. In: Proc. of the Fourth IFCIS Intl. Conf. on Cooperative Information Systems, Edinburgh, Scotland, Sep. 1999
23 Snoussi H, Magnin L, Nie J-Y. Toward an Ontology-based Web Data Extraction, 2002. http://www.cs.unb.ca/ai2002/baseweb/BASeWEB2002_Paper3.pdf
24 Embley D W, et al. A Conceptual-Modeling Approach to *Extr*acting Data from the Web. In: Proc. of the 17th Intl. Conf. on Conceptual Modeling (ER'98), Singapore, Nov. 1998. 78～91. http://faculty.cs.byu.edu/～dennis/papers/er98.ps

(下转第151页)

法的应用范围。在跟踪性能上，模型集和 IMM 均自适应的算法优于 AGIMM 算法，但两者都有较大的峰值误差。进一步试验表明这两种方法的性能不仅与目标机动性有关还与常量有关。

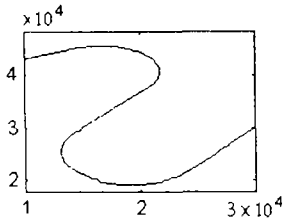

图1 目标运动轨迹

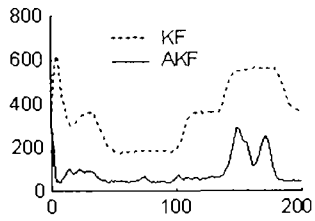

图2 $F(w=-5.6)$均方误差曲线

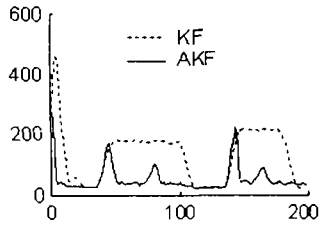

图3 $F(w=0)$均方误差曲线

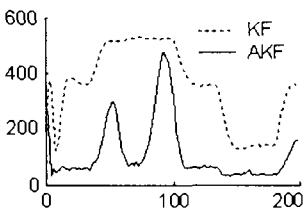

图4 $F(w=5.6)$均方误差曲线

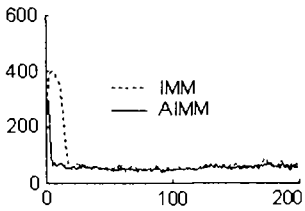

图5 位置均方误差曲线

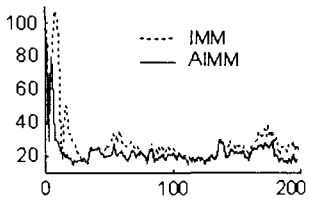

图6 速度均方误差曲线

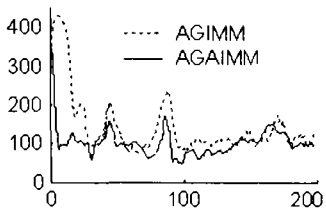

图7 位置均方误差曲线

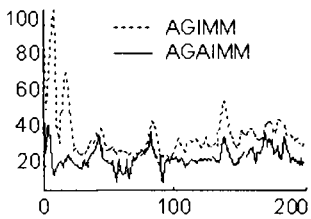

图8 速度均方误差曲线

参 考 文 献

1 Mazor E, Averbuch A, Bar-Shalom Y. Interacting Multiple Model Methods in Target Tracking: A Survey. IEEE Transactions on Aerospace and Electronic Systems, 1998, 34(1)

2 张安民，杨世兴，李志舜. 目标跟踪中的混合多模型方法：综述. 系统工程与电子技术，2001，23(4)

3 Campo L, Mookerjee P, Bar-Shalom Y. State Estimation for Systems with Sojourn-Time- Dependent Markov model Switching. IEEE Trans. on Auto. Con., 1991, 36(4)

4 Li X R, Bar-Shalom Y. Multiple Model Estimation with Variable Structure. IEEE Transactions on Automatic Control, 1996, 41(4)

5 Efe M, Atherton D P. Maneuvering Target Tracking Using Adaptive Turn Rate Models in The Interacting Model Algorithm. In: Proc. of the 35th Conf. on Decision &Control, 1996. 3151～3156

6 Jilkov V P, Angelova D S. Design and Comparison of Mode-Set Adaptive IMM Algorithms for Maneuvering Target Trac-king. IEEE Transactions on Aerospace and Electronic Systems, 1999, 35(1)

7 周宏仁，敬忠良，王培德. 机动目标跟踪. 国防工业出版社，1991

（上接第131页）

25 Atzeni P, Mecca G. Cut and Paste. In: Proc. of the PODS'97, 1997

26 Gupta A, Harinarayan V, Rajaraman A. Virtual database technology. SIGMOD Record, 1997, 26(4): 57～61

27 Hammer J, Garcia-Molina H, Cho J, Hranha R, Grespo. Extracting semistructured information from the web. In: Proc. of the Workshop on Management of Semistructured Data. May 1997

28 http://www.dia.uniroma3.it/Araneus/

29 http://www.deg.byu.edu/

30 http://abrolhos.lbd.dcc.ufmg.br/~debye/

31 http://www.delphi-jedi.org/

32 Adelberg B. Nodose: A tool for semiautomatically extracting structured and semistructured data from text documents. In: Proc. of ACM SIGMOD Conf. on Management of Data, 1998

33 http://www.rr.com/rdrun/

34 http://research.compaq.com/SRC/WebL/

35 http://www.cse.ogi.edu/sysl/projects/XWRAP/xwrap.html

36 http://www.atg.com/en/products/portalsuite/web_application_gear.jhtml

37 http://www.downlinx.com/proghtml/299/29907.htm

38 http://www.x-fetch.com/xhtml/wrapper.html

39 http://cheops.cis.upenn.edu/W4F/

40 http://www.wisosoft.com/

41 Laender A H F, Ribeiro-Neto B, da Silva A S, Teixeira J S. A Brief Survey of Web Data Extraction Tools. ACM SIGMOD Record, 2002, 31(2)

42 Kuhlins S, Tredwell R. Toolkits for Generating Wrappers - A Survey of Software Toolkits for Automated Data Extraction from Websites. http://citeseer.nj.nec.com/539694.html

43 BrightPlanet.com DeepWeb White Paper. http://www.completeplanet.com/Tutorials/DeepWeb/index.asp

44 Yau S T. Extracting Data Behind Web Forms, Technical Report, Brigham Young University, Provo, Feb. 2002. http://blon-die.cs.byu.edu/CS652/DataBehindForms.pdf

45 http://www.w3.org/People/Raggett/tidy/

46 http://www.republica.com/

47 http://www.w3.org/TR/2001/NOTE-data-extraction-20011031

48 http://www.w3.org/2001/sw/