第31卷 第20期 计 算 机 工 程 2005年10月
Vol.31 № 20 Computer Engineering October 2005
·基金项目论文· 文章编号：1000—3428(2005)20—0054—02 文献标识码：A 中图分类号：TP301.6

基于网页结构树的 Web 信息抽取方法

陈 琼，苏文健
（华南理工大学计算机科学与工程学院，广州 510640）

摘 要：提出了网页结构树提取算法及基于网页结构树的 Web 信息抽取方法。抽取信息时，在网页结构树中定位模式库中的待抽取信息，用模式库中的待抽取信息和网页结构树的叶结点对应的网页信息进行匹配。因而对网页信息的抽取，可以转化为对网页结构树的树叶结点信息的查找。实验证明，该方法具有较强的网页信息抽取能力。

关键词：信息抽取；半结构；网页结构树；模式

Web Information Extraction Based on Web Structure Tree

CHEN Qiong, SU Wenjian
(School of Computer Science and Engineering, South China University of Technology, Guangzhou 510640)

【Abstract】This paper proposes an algorithm that is used to construct the Web structure tree and a Web information extraction method based on Web page structure tree. While extracting information, it locates the information that should be extracted in the Web page structure tree and matches the pattern information with the terminal information in Web page structure tree. The Web information extraction is the terminal information extraction in Web page structure tree. This method can efficiently extract information from Web pages.

【Key words】Information extraction; Semi-structure; Web page structure tree; Pattern

网页信息抽取技术，包括基于归纳学习的信息抽取、基于 HTML 结构解析的信息抽取、基于 Web 查询的信息抽取、基于自然语言处理的信息抽取、基于模型的信息抽取和基于本体的信息抽取[1]。其中基于 HTML 结构解析的信息抽取的特点是，将 Web 文档转换成反映 HTML 文件层次结构的解析树，通过自动或半自动的方式产生抽取规则。典型的系统有 W4F[2]、RoadRunner[3]、XWRAP[4]等。

本文提出的基于网页结构树的 Web 信息抽取方法，利用网页结构树提取算法，构造网页结构树，使得页结点包含网页内容信息。抽取信息时，在网页结构树中定位模式库中的待抽取信息，用模式库中的待抽取信息和网页结构树的叶结点对应的网页信息进行匹配。因而对网页信息的抽取，可以转化为对网页结构树的树叶结点信息的查找。

1 网页结构树的构造

1.1 HTML 文档特点

目前，Web 上的数据大部分是以 HTML 的形式出现的。HTML 文档由标记(TAG)和元素组成。HTML 标记确定了浏览器所显示文档元素的格式，大多数 HTML 标记是成对出现的，它们分别用作开始标记和结束标记，HTML 的结束标记与开始标记的唯一区别是多了个斜杠“/”。HTML 标记放在尖括号里，如 <HTML> 是位于 HTML 文档中的第一个条目，

HTML 文档由标题<HEAD>和主体<BODY>两部分组成。标题部分包含文档的标题，主题部分包含文档的内容。为标明 HTML 文档标题部分的起始和结束，可以使用标记<HEAD>和</HEAD>；同样，可以使用标记<END>和</END>标明文档主体部分的开始和结束。<TITLE>和</TITLE>用于显示 WEB 页面的标题。<TABLE>，</TABLE>是表格起始和结束标记，<TH>，</TH>是表头起始和结束标记，<TR>，</TR>是表行起始和结束标记，<TD>，</TD>是表格内容起始和结束标记等。

1.2 网页结构树的构造算法

利用 HTML 的 TAG 的特征，采用标记匹配和回溯相结合的方法构造 Web 文档结构树。大多数 HTML 标记是成对出现的，在起始标记和结束标记之间，包括网页描述属性信息和网页内容信息，如 <td width = "33%">型号：</td>。在起始标记<td>和结束标记</td>之间的 width ="33% " >是属性信息，“型号：”是网页内容信息。在构造网页文档结构树时，忽略属性描述信息，因此只需对部分 TAG 标记进行分析。如果主要是对表格内容的抽取，则需要考虑的 HTML 标记主要有 <HTML>, </HTML>, <BODY>, </BODY>,<TABLE>,</TABLE>,<TH>,</TH>,<TR>,</TR>, <TD>, </TD>, <IMG>。对于其它的 HTML 标记可视为无用 HTML 标记，在程序处理中将忽略对这些标记的处理。

网页文档结构树的每个结点对应一个 Tag 标记。因此构建 TagNode 树的前提条件是正确地读取标记，分析开始标记、结束标记和没有得到匹配的标记。结点对应的 Tag 开始与结束标记之间的内容存在 TagNode 类成员 data 中。

网页结构树构造算法 TagNode 如下：

如果读取的文件没有到文件尾，作以下操作：

如果获取标记成功

基金项目：国家自然科学基金资助项目（60003019）；广东省自然科学基金资助项目（990582）；广东省科技攻关资助项目项目（C10201）
作者简介：陈 琼（1966－），女，副教授，主研方向：机器学习，智能信息计算；苏文健，硕士生
收稿日期：2004-10-04 **E-mail：**qiongchen66@yahoo.com

如果为开始标记且根结点为空
创建根结点，使当前结点为根结点
如果为开始标记且根结点不为空
如果获取的标记为”img”
根据获取标记创建新结点，使之成为当前结点的儿子
标记该新创建的结点为匹配结点
如果获取的标记和当前结点不同
根据获取标记创建新结点，使之成为当前结点的儿子，使新创建的结点为当前结点
获取当前结点的内容
如果获取的标记和当前结点相同
根据获取标记创建新结点，使之成为当前结点的儿子，标记当前结点为匹配结点，并使新创建的结点为当前结点
获取当前结点的内容
如果为结束标记
如果为当前结点标记的匹配结束标记
标记该结点为匹配结点使当前结点为其父结点
如果没有找到和该结束标记匹配的结点，作以下操作：
回溯到当前结点的第一个未匹配的前辈结点
如果是和结束标记匹配的结点
标记该前辈结点为匹配结点，使其父结点为当前结点

通过输入一个网页，例如图 1 所示的网页。

图 1　输入网页

TagNode 算法可以自动地对网页结构进行分析，构造其 TagNode 树，如图 2 所示。

图 2 网页结构树（部分）

图 2 只是网页结构树的一部分，网页的内容信息都对应在树结点上，叶结点对应网页的最小内容信息单元。TagNode 算法构建的网页结构树虽然没有反映网页的全部信息，但通过网页结构树，用户可以了解网页的结构，还可以查看网页中他们感兴趣的信息，从而对网页进行信息抽取。

2 网页信息的抽取

2.1 网页信息抽取算法

利用 TagNode 算法构造的网页结构树，可以把网页信息的抽取映射为在网页结构树中信息的查找。这里设计了抽取手机信息的网页信息抽取器。通过对不同网站的手机信息网页的分析，发现具有这样的规律：要抽取的手机型号和手机价格通常在同一个 table 中，并且手机型号和手机价格间存在一对一或一对多的关系。因而可采用如下启发式规则进行网页信息抽取。启发式规则为：待抽取的信息的各部分通常在同一个 table 中，并且它们之间存在一对一或一对多的关系。

在进行信息抽取前，首先建立模式库。模式库包含待抽取信息的表述、特征项等。抽取信息时，在网页结构树中定位模式库中的待抽取信息，用模式库中的待抽取信息和网页结构树的叶结点对应的网页信息进行匹配，如果匹配成功，则找到了一部分要抽取的信息。利用上述启发式规则，其他待抽取信息对应的结点和这个已匹配的叶结点在同一个 table 中。因此，模式库中的待抽取信息或特征项只需和已匹配的树叶结点的兄弟结点对应的网页信息比较，通常就可匹配成功，完成信息的抽取。

网页信息抽取算法如下：

如果要处理的网页结构树结点的儿子为 NULL 及该结点标记的字符串的前两个字符为“ td”
如果 dataKind 的值为 UnKnown 或 Cost
根据学习的模式修改 regularText
匹配待抽取信息
如果匹配成功
matchString 记录下匹配的信息
找出含匹配信息的结点所在的最底层的“ table ”父结点
matchLocation 赋值为对该“ table ”父结点的引用
dataKind 赋值为 Kind
如果 dataKind 的值为 Kind
匹配其他待抽取信息
如果匹配成功
如果含有匹配信息的结点所处的最底层的“ table ”父结点与 matchLocation 相同
抽取信息并存入数据库

此算法中，要注意几个全局变量的意义：

dataKind：记录抽取数据时的当前数据类型

matchLocation：用于指示抽取数据时所抽取树结点所处于的最底层的"table"父结点

matchString：用于存储抽取数据时所抽取的信息

regularText：用于抽取手机数据的正则表达式字符串

2.2 抽取器模式学习算法

通过使用网页信息抽取器，可以对输入的网页进行信息抽取。固定不变的模式库将限制网页信息抽取器的灵活性。抽取器必须具有学习能力，能够通过不断地学习去抽取模式、丰富模式，才能抽取更多的信息，具有更广泛的用途。

考虑到模式学习的需要，设计了人工学习和自动学习的两种方式，补充和丰富网页信息抽取器的模式库，从而提高抽取器的灵活性。

(下转第 140 页)

关，并进行了对比测试，有 4 组数据：(1)IPv4 局域网中两台 IPv4 主机直接通信，表示为 v4-v4；(2)IPv6 局域网中两台 IPv6 主机直接通信，表示为 v6-v6；(3)IPv4 局中两台 IPv4 主机通过 Windows 2000 中的路由转发功能进行通信，表示为转发；(4)一台 IPv6 主机和一台 IPv4 主机通过转换网关进行通信，表示为网关。

3.1 时延测试

时延测试使用 Ping 命令实现，Ping 包的大小为 1 024B，然后取 100 个 Ping 包的平均时延，测试结果如图 3 所示，从图中可以看出，转换网关的时延只比 Windows 2000 的路由转发多花了大约 25ms，表明转换网关的时延是非常低的。

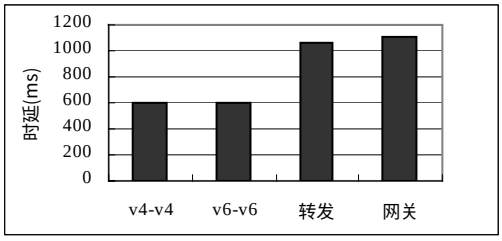

图 3 时延测试

3.2 带宽测试

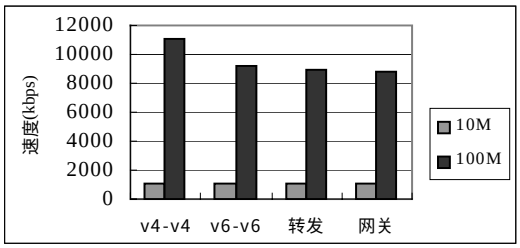

图 4 带宽测试

带宽测试使用 ttcpw 程序，分别在 10Mbps 和 100Mbps 网络中发送 100MB 数据，测试的结果如图 4 所示。从图 4 中可以看出，对于 10Mbps 网络，4 种测试环境的的带宽基本是一样的，对 100Mbps 网络，转换网关和路由转发的带宽都有一定的下降，但还是能包达到 9 000kbps。比较转换网关和路由转发的带宽，转换网关的带宽略有下降，大约为 50kbps。

3.3 应用程序测试

我们对转换网关进行了 FTP 协议、HTTP 协议、Telnet 协议和 DNS 的测试，转换网关顺利地通过了这些测试，并取得了良好的测试效果。

4 结论

本文提出了一个基于 Windows 操作系统的全功能转换网关架构，并进行了代码实现，然后对转换网关进行了时延、带宽和应用程序测试，取得了满意的测试结果，表明这是一个优秀的转换网关解决方案。本系统虽然是基于 Windows 操作系统提出的，但对其他的系统也具有很大的借鉴意义。

参考文献

1 Srisuresh T,Egevang K.Traditional IP Network Address Translator (Traditional NAT) [S]. RFC 3022,1995-12

2 Mordmark E.Stateless IP/ICMP Translation Algorithm (SIIT) [S]. RFC 2765,2001-01

3 Tsirtsis G,Srisuresh T.Network Address Translation—Protocol Translation (NAT-PT) [S]. RFC 2766,2000-02

4 Thomson S, Huitema C. DNS Extensions to Support IP Version 6[S]. RFC 1886,2000-02

5 Srisuresh T,Tsirtsis G. DNS Extensions to Network Address Translators (DNS_ALG) [S]. RFC 2694, 1999-09

6 美国微软公司.Windows 2000 驱动程序开发大全(第 1 卷 设计指南) [M]. 北京:机械工业出版社, 2001-08

(上接第 55 页)

(1)人工方式。打开一个网页后，用户尝试抽取网页中的信息，但是用户需要的模式并不存在。此时，用户可以选择要抽取的信息，并把它添加到模式库中。这样，日后用户再碰到同类的信息后，网页信息抽取器就可以实现对该类信息的自动抽取。可以采用选中网页结构树中的某个结点或选择文本框的字符加入模式库。

(2)自动方式。当用户需要的模式不存在的时候，用户可以输入大量的样本网页让抽取器进行学习。通过对大量网页样例的学习，把原来模式中没有的待抽取信息表述通过学习而添加到模式字符串中。这样，抽取器就可以通过不断的学习而抽取不同的网页了。

3 结果分析及未来的工作

使用网页信息抽取器对十多个手机网站的网页进行分析，网页信息抽取器能够为每个网页构造结构树，对于特定网站的所有具有相似结构的网页信息，都可以正确抽取，对于不同结构的网页，通过模式学习，可以正确抽取用户感兴趣的的信息。

这里的网页结构树构造算法主要处理的是表格标记，可以推广到处理“ Li ”,“ Ol ”,“ Ul ”“ Hx ” 等标记。考虑属性信息，可以构造更全面反映网页结构及特征的网页结构树。

另外，通过模式的学习，本网页信息抽取器的应用领域可以推广到其他领域。对于其它领域的信息，通过模式的学习，可以把领域知识存入模式库，从而使信息抽取器具有更强的灵活性。

参考文献

1 Laender H F, Ribeiro-Neto B A, A S da Silva, et al. A Brief Survey of Web Data Extraction Tools. SIGMOD Record, 2002, 31(2): 84-93

2 Sahuguet A, Azavan F. Building Intelligent Web Applications Using Lightweight Wrappers. Data and Knowledge Engineering, 2001,36 (3), 283-316

3 Crescenzi V, Mecca G, Merialdo P. RoadRunner: Towards Automatic Data Extraction from Large Web Sites. Rome, Italy: In: Proceeding of the 26th International Conference on Very Large Database Systems, 2001:109-118

4 Liu L, Pu C, Han W. XWRAP: An XML-enable Wrapper Construction System for Web Information Sources. San Diego, California: In: Proceedings of the 16th IEEE International Conference on Data Engineering, 2000: 611-621

5 李 晶，陈恩红. Web 信息抽取. 计算机科学, 2003, 30(6):78-81