李广建(北京大学信息管理系，北京，100871)
乔建忠(中国科学院国家科学图书馆，北京，100190；中国科学院研究生院，北京，100049；解放军艺术学院教育技术中心，北京，100081)

全自动生成网页信息抽取包装器的主要技术方法研究

Research on Major Technical Methods of Fully Automatic Wrapper Generation for Web Information Extraction

[摘要]网页信息抽取包装器的生成方法很多，按自动化程度可分为手工、半自动和全自动三类，本文旨在研究全自动生成网页信息抽取包装器的主要技术方法，首先构建了对应的分类体系，其次对近年来主流的十五种包装器生成技术进行了定性分析和分类比较，最后提出五点发展趋势。

[Abstract] Many methods to address the problems of wrapper generation for web information extraction have been proposed in the literature in the last few years. Based on the degree of automation, they can be divided into three categories: manual, semi-automatic and fully automatic. This paper aims to study fully automatic wrapper generation methods of web information extraction. Firstly, a fully automated wrapper generation technology classification framework is constructed. Secondly, 15 major fully automatic wrapper generation technologies in

recent years are compared, analyzed qualitatively and summarized. Thirdly, the future development trends are summed up.

[关键词] 网页信息抽取；包装器；包装器生成技术；结构化信息；深层网

[Keywords] Web Information Extraction; Wrapper; Wrapper Generation Technology; Structured Information; Deep Web

[分类号] G250.73

1 引言

网页信息抽取是信息抽取中的一类，网页信息抽取的包装器生成技术目前发展为一个较为独立的领域。包装器是由一系列抽取规则以及应用这些规则的计算机代码组成的，专门从特定信息源中抽取需要信息并返回结果的程序[1]。全自动包装器生成技术是以全自动方法生成抽取规则的技术。该技术的优点主要有两个：一是加快了处理网上海量信息的速度，提高了获取深层网(Deep Web)信息的能力；二是把不同信息源的信息整合到一个数据库中，方便比较和分析利用，如比价和数据挖掘等。

Line Eikvil 在文献[1]中对网页信息抽取技术进行了全面总结，同时比较了 1996 年至 1998 年出现的七个包装器生成

技术：ShopBot、WIEN、SoftMealy、STALKER、RAPIER、SRV 和 WHISK，并把它们分为处理结构化较强和较弱两类网页信息抽取包装器。七个工具均采用人工或半自动化包装器生成技术，这与本文的十五种全自动包装器生成技术有所不同。

Alberto H. F. Laender 和 Altigran S. da Silva 等人在文献[2]中把抽取各类信息的工具分为六类，并比较了 2002 年以前出现的十五个信息抽取工具，其中抽取网页信息的包装器生成技术只是作为信息抽取工具所采纳技术中的一类，并且全自动信息抽取工具只有两个：RoadRunner 和 XWRAP，而本文主要研究全自动包装器生成技术的分类方法以及对 2001 年以后的十五种全自动包装器生成技术的定性分析与比较。

本文在第二部分构建了一个全自动包装器生成技术的分类体系，在此基础上第三部分对十五种全自动包装器生成技术进行了定性分析和分类比较，最后总结了该领域今后的发展趋势。

2 全自动包装器生成技术的分类

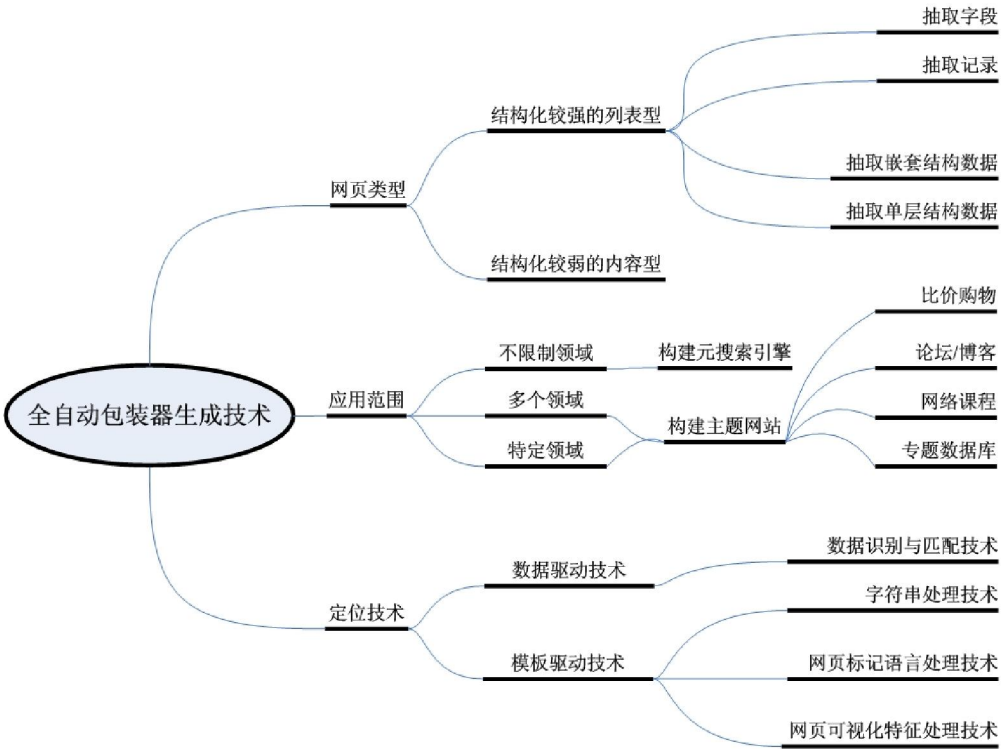

图 1：全自动包装器生成技术分类体系示意图

目前被称为是全自动的包装器生成技术主要有 RoadRunner[3]、SG-WRAP[4]、EXALG[5]、MDR[6]、DeLa[7]、Xpath-Wrapper[8]、DEPTA[9,10]、ViPER[11]、ViNTs[12]、ViDRE[13]、WDE[14]、PSTN[15]、Grubber[16]以及未给出包装器生成技术英文缩写名称的文献[17]和文献[18]，它们是在 2001 年以后相继出现且他引率较高。本文在研究其技术特点和共同特征的基础上按照网页类型、应用范围和定位技术三个标准对全自动包装器生成技术进行了分类并粗略地构建了一个分类体系框架，如图 1 所示。

抽取信息的复杂程度主要取决于网页结构化的强弱，而网

页的结构化程度则取决于用户想要抽取的数据属性。图中列表型网页代表具有严格结构化的网页，内容型网页代表松散的结构化网页。在处理列表型网页上如果抽取的数据属性不同又有抽取列表中记录和抽取记录中字段之分；如果列表结构的复杂程度不同又有处理多层嵌套结构和处理单层结构之分。

全自动包装器生成技术的应用范围面向何种领域主要取决于抽取信息的动机。面向特定领域的全自动生成技术专用于构建某个特定主题网站或专题数据库。面向多个领域是用在多个特定主题领域，实用性更强。不限制领域一般用于构建元搜索引擎，此类技术用于信息抽取的目的并不限于某个或某几个特定领域，因此通用性较强。

所谓定位技术，就是准确识别出目标数据的技术，它是决定如何生成抽取规则所采用的算法，也是包装器生成的核心技术。模板驱动技术是目前主流的包装器生成技术，它以准确还原由数据库动态生成网页的“假设模板”为基础，从而抽取出嵌在模板中的数据。与传统的模板依赖和参数假设不同的是数据驱动技术，其应用语义描述和实体识别技术直接识别和匹配目标数据来生成抽取规则。定位技术在各包装器中的运用在下一节有具体论述。

3 全自动包装器生成技术的定性分析与比较

3.1 属性定义与比较

本文把对复杂数据的支持、定位技术的运用、多种网页语言的支持以及人工干预的程度作为对全自动包装器生成技术进行定性分析与比较的四个属性。

对复杂数据的支持，这一属性具体表现为以下七种能力：A.能处理严格的结构化网页；B.能处理松散的结构化网页；C.能处理多层嵌套列表结构网页；D.能处理单层列表结构网页；E.不受数据区域数限制；F.受数据区域数限制(F4 表示只能处理四个以内数据区域)；G.能抽取列表中的字段属性。

定位技术的运用，这一属性具体表现为以下四个方面：A.基于网页标记语言的处理技术；B.基于网页可视化特征的处理技术；C.单纯基于网页字符串的处理技术；D.基于数据识别与匹配的数据驱动技术。其中 A、B 和 C 代表了模板驱动技术，D 代表了数据驱动技术。

对多种网页语言的支持，这一属性具体表现为以下三个方面：A.与网页语言无关；B.支持 XML；C.支持 HTML。

人工干预的程度，这一属性体现了自动化的程度，即使是全自动包装器生成技术也有少量的人工指导，具体表现为以下四种情况：A.人工指定抽取模式；B.人工校正抽取规则；C.人工生成正或反训练样本；D.无人工干预。

如果上述四个属性下的具体表现用其前面的字母代替，那么十五种全自动包装器生成技术的定性比较结果如表 1 所示：

属性 / 内容	复杂数据	定位技术	网页语言	人工干预	年份
Grubber	A、B、C、G	A、B	C	D	2008 年
PSTN	A、B、F4	A、B	C	D	2008 年
文献[17]的技术	A、B	D	C	D	2008 年
文献[18]的技术	A、C、G	C	C	D	2007 年
WDE	A、B	A	C	B	2007 年
ViDRE	A	B	A	B	2006 年
ViNTs	A、C、E	A、B	C	D	2005 年
ViPER	A、E	A、B	C	D	2005 年
DEPTA	A、C、E、G	A、B	C	D	2005 年
XPath-Wrapper	A、B、G	A	B、C	C	2005 年
DeLa	A、C、G	C	C	D	2003 年
MDR	A、E	A	C	D	2003 年
EXALG	A、B、C	A	C	D	2003 年
SG-WRAP	A、B	A	C	A、B	2002 年
RoadRunner	A、B、C	A	C	D	2001 年

表 1：15 种全自动包装器生成技术的定性比较表

3.2 定性分析与特点概述

3.2.1 对复杂数据的支持

对嵌套结构网页的处理能力是支持复杂数据抽取包装器生成技术的重要特征。所谓嵌套结构就是列表型网页中目标数据行仍由数据列表构成，亦称为多层结构。RoadRunner、DeLa 和文献[18]提出的包装器生成技术就把对嵌套结构数据的处理作为自己的一个突出特点。

RoadRunner 首次提出“不匹配”的目标数据定位技术，为此需要一次同时处理两个网页，其主要设计思想是免集合运

算符正则表达式(Union-Free Regular Expressions，UFRE)与嵌套类型有着密切的联系。UFRE 由 NULL、#PCDATA 和字符集Σ组成，其中#PCDATA 表示可被解析的嵌套元素类型。嵌套类型τ由归纳出的代表 UFRE 的 σ 来计算得到，其计算公式为 τ = type(σ)。这种通过给定网页中的字符串元素用免集合运算符正则表达式 σ 来逆向推导源数据的过程被作者称为模式发现，其关键是找到最小化 UFRE。对于两个 UFRE 代表的字符串集合 $L(\sigma_1)$ 和 $L(\sigma_2)$，如果 $L(\sigma_2)$ 包含 $L(\sigma_1)$，那么 $L(\sigma_1)$ 和 $L(\sigma_2)$ 的最小化 UFRE 就是字符串集合的至少上界 (the Least Upper Bound of UFREs)，用公式 $LUB(\sigma_1, \sigma_2)$ 表示。而用 UFRE 来抽象表示多样的网页嵌套类型仍然具有局限性，因此该技术更适合处理结构化相对较强的网页。

DeLa 处理嵌套结构的方法是利用 DOM 树的等级模式来发现嵌套结构，这是一种基于后缀树和字符串比对来寻找连续重复模式的算法，其算法思想分成两步完成。第一步，抽取出目标数据区。为实现此任务，Dela 的设计者提出了“富数据区抽取”的概念。其含义是一次比较两个网页，为这两个网页分别建立 DOM 树，按“深度优先”的顺序，对两个 DOM 树逐个节点进行比较，剔除那些在相同位置重复出现的分支(可能是广告等干扰性数据)。第二步，将网页代码定义为一个大字符串，通过后缀树算法经过一定的迭代过程找出连续重复模式即网页中的相邻重复子串。此模式并不专指用户感兴趣的数据部

分，而是任何网页字符串。为进一步剔除干扰，从候选重复模式中确定正确的连续并且重复的模式需依照三条规则：一是忽略包含兄弟模式的模式；二是如果无兄弟模式的情况下忽略包含父模式的模式；三是选择覆盖面最大的重复模式，覆盖面等于字符串长度与其出现次数的积。网页中的嵌套结构表现为父连续重复模式嵌套子连续重复模式。子连续重复模式出现的次数要么是 0 要么是多次。如果一个子连续重复模式用一个“*”来标记，那么生成的包装器用免集合运算符正则表达式表示为“<P>(<A>text</A>)*text</P>”或“(AB*A)*C*”。表达式中的“*”表示其前面是一个子连续重复模式。

文献[18]采用与 DeLa 基本相同的方法来识别嵌套结构。他们所使用的方法只适用于处理拥有高级检索功能的基于后台数据库动态生成的搜索结果页这种结构化较强的网页类型。而且 DeLa 和文献[18]处理嵌套结构的设计思想都使用了后缀树算法，因为该算法在线性时间内能较好地解决字符串查找和匹配问题，快速发现和定位字符串中的重复子串。

3.2.2 定位技术的运用

依图 1 所示的分类体系，下面笔者将从 MDR 和 DEPTA 入手，概述各包装器生成技术在运用定位技术上的主要特点和变化。需要说明的是 DeLa 和文献[18]采用的定位技术是单纯基于网页字符串处理技术的，其特点已在前文概述本节不再重复。

3.2.2.1 以网页标记语言处理技术为基础

MDR 和 DEPTA 是出自美国芝加哥伊利诺斯大学计算机科学系的两个包装器系统。DEPTA 在 MDR 的基础上进行了改进和创新。MDR 是单纯基于网页标记语言的处理技术来定位和抽取数据的，而 DEPTA 则加入了网页可视化特征的处理技术。

MDR 所使用的定位技术是基于作者对购物网站网页特征的两点观察：一是某个网页可能有多个数据区，每个数据区由数据记录组成，各数据区有着相同的网页标记符号，可通过字符串匹配算法找出网页中的相同标记；二是网页代码中的标记符号自然形成一棵标记树，一组相似的数据记录必定在同一个父节点下，并表现为这个父节点下的一组子树，而且同一条数据记录不会横跨并列的两个子树。关键问题是如何找到这个父节点，具体实现分三步完成。首先，建立网页的标记符号树；其次，利用字符串比对的方法找到网页中的目标数据区；最后，识别出目标数据区中的数据记录。字符串和标记树的比对实际上是比较各节点的编辑距离值。对数据记录的识别，MDR 是基于某种假设，就是同属于一个数据区的记录具有相同的网页标记结构。显然这种假设没有考虑嵌套结构的情况。

与 MDR 采用相似定位技术的包装器生成技术还有 SG-WRAP、EXALG、XPath-Wrapper 和 WDE。

SG-WARP 虽然也基于网页标记语言定位技术抽取目标数据但它增加了数据语义描述--DTD 文档的辅助。其实现策略是首

先解析网页建立 DOM 树，用 DOM 树作为网页的结构描述，其中的目标数据项是树中的叶节点，目标数据的位置信息用叶节点到根节点的路径来描述；然后由用户指定抽取目标数据的描述，该数据描述被用来定义成 DTD 文档。DTD 文档作为网页的语义描述，系统会自动生成目标数据的结构描述与语义描述的匹配规则，并从中抽取出目标数据。DTD 文档的最终目的是用来生成 XML 文件。SG-WARP 在自动生成包装器之前和之中都专门设计了人工干预的功能，即由人工来指定抽取模式和校正匹配规则。

EXALG 与 RoadRunner 的定位技术类似，但做了针对性的改进。EXALG 假设来源网页都是由后台数据库按照一定的模板生成，如果从网页能正确推导出模板，那么按照模板就能准确抽出网页中的目标数据。对于如何推导模板，EXALG 与 RoadRunner 的处理技术不同，EXALG 并不赞成用正则表达式来表示抽取规则而是使用自己定义的元组函数与集合函数来表示网页模板。对于网页标记或目标数据也用专门的类型符号表示。这些类型符号就是元组函数和集合函数中的变量。元组函数代表一片数据区域或整个网页，集合函数代表数据区内的数据记录，数据记录中的不同字段使用类型符号表示。针对可选数据项和网页标记语言符号出现在目标数据中的情况，EXALG 也做了专门设计。EXALG 与其他基于网页标记语言的处理技术类似也使用了树编辑距离来计算节点的位置，并通过计算类型

符号的发现频率和定义不同的角色来最终推导出表示模板的公式。

XPath-Wrapper 所使用的定位技术是典型的基于网页标记语言处理技术的包装器生成方法。其基本思想是首先解析 HTML 或 XML 网页为 DOM 树，然后借用 W3C 标准--XPath 语言遍历该 DOM 树，使用事先学习到的 XPath 表达式来定位目标数据。不过 XPath-Wrapper 介入了一定的人工干预。它先由人工生成正、反训练样本，再通过遍历训练样本学习到较为准确的 XPath 抽取规则。XPath-Wrapper 专用于抽取基于 XSLT 的网页信息如 XML 文档或从 XML 经 XSLT 转换后的 HTML 网页。

WDE 同样是使用了基于网页标记语言的定位技术，其突出特点是针对结构化较弱的网页类型特征，在计算节点的树编辑距离值时加入了权重因子对节点的评价，即内部节点权重较高，因为它们更接近目标数据。WDE 采用的关键步骤先是解析网页生成网页标记树；其次，计算节点的树编辑距离值按节点的相似性将节点聚类；第三，计算出候选数据记录，并按相似性将数据记录聚类，由此生成数据抽取模板。另外 WDE 也介入了一定的人工干预，即从系统生成的多个抽取规则中选出最佳抽取规则，一般应用此规则抽取的数据记录最多。

DEPTA 针对单纯依赖网页语言处理技术的 MDR 做了三点改进：一是借助网页的可视化特征建立网页的 DOM 树，因为看上去工整的网页其网页标记可能是不合语法的，所以单纯利用网

页标记语言建立 DOM 树会造成信息的缺失；二是在抽取数据记录的基础上增加了抽取数据字段的功能，为此，其提出“局部树比对算法”；三是利用局部树比对算法实现了对嵌套结构数据的抽取。尽管 DEPTA 对网页可视化特征处理技术的应用还较初步，但这种结合代表了后来包装器生成的新趋势。

3.2.2.2 可视化特征处理技术的正式介入

与 DEPTA 利用网页可视化特征的处理技术不同，ViPER、ViNTs、ViDRE、PSTN 和 Grubber 将网页可视化特征直接用于目标数据的定位。利用网页的可视化特征来定位目标数据是为了弥补单纯依赖网页标记语言的一些不足，诸如网页标记语言的语义表达较弱、语法结构较随意、网页标记符号可能作为目标数据的组成内容等。

ViPER 同时基于网页标记语言和网页可视化特征处理技术来定位目标数据。在网页标记语言处理技术上它借鉴了 MDR 的做法，而网页的可视化特征被用来分割数据区并为之分配权重。分割数据区时 ViPER 同时考虑了横向与纵向的视觉特征，这对目标数据区按左右结构排列时的正确抽取比较有效。为网页中不同数据区分配权重时，ViPER 考虑的是数据区的位置和面积大小，一个位置居中，面积较大的数据区权重也较大，这比较符合大多数搜索结果区在网页中的表现特征。

ViNTs 利用网页可视化特征处理技术的特点是针对搜索引

擎结果页面的特征定义了数据行和数据块的概念，数据块由数据行组成。除了目标数据块以外，网页中还存在一些诸如广告之类的干扰数据块。ViNTs 定义了数据块的位置、类型和形状三个可视化特征变量，并将相似的数据块聚类。在识别出目标数据块的基础上，再应用基于网页标记语言的处理技术生成包装器正则表达式。ViNTs 与 ViPER 不同的是它没有考虑数据块的纵向分割问题。

由于网页标记语言的多样性，ViDRE 考虑到基于某种网页标记语言来定位目标数据存在较大的局限性，因而设计了一种与网页语言无关的单纯依赖网页的可视化特征来定位目标数据的方法。其技术特点是考虑了搜索结果页面的四项可视化特征，即目标数据区的位置特征、数据记录的排列特征、形状特征和内容特征。根据四项特征利用 VIPs(Vision-based Page Segmentation)算法[19]生成网页的数据块可视化特征树。先根据位置特征从视觉块树中定位目标数据区，再依据排列和形状特征从视觉块树中抽取数据记录。

PSTN 的特点是未使用树编辑距离，而是运用信息所在的结构体之间存在的某种关系来实现对信息的抽取。结构体被定义为网页标签树中的一棵子树。网页的可视化特征被用来判断结构体之间的关系。如通过对网页中各数据区分布特征的分析，可得出结构体等势原理，即同一数据区域内的各条数据所在的最小结构体拥有相同的结构体嵌套关系。结构体的嵌套关

系可用标签路径来表示。PSTN 根据结构体等势原理提出结构化分离算法，并最终定位每一条目标数据记录。只是 PSTN 对数据区域数的识别限制在四个以内。

Grubber 综合运用网页标记语言处理技术和网页可视化特征处理技术来提出一种识别和表示网页模板的方法。Grubber 定义了三类网页标记：开始、文本和结束标记。其中开始标记的属性中就包含了可视化特征属性(如表格的 Left、Top、Height 和 Width 值)。Grubber 借助统计方法、分析网页结构和网页布局特征等手段生成聚类标记—CTokens，并用联合正则表达式(Union Regular Expressions，URE)表示抽取规则。Grubber 基于聚类标记推导模板的方法仍然是一种基于正例的正规语法推导问题，因此在理论上其正确性并不强，EXALG 对此已进行改进。

3.2.2.3 数据驱动的定位技术

文献[17]所使用的定位技术不同于传统的模板驱动、模板依赖的方法，而是一种数据驱动的包装器生成和维护的方法。文献[17]的核心思想是定义了语义块的概念来定位目标数据。语义块由网页中数据项的位置信息、语义信息以及值构成的三元组表示。语义块被分为源语义块和目标语义块两类，分别代表了正样本或改版前网页中的数据块和待抽取页面或改版后网页中的数据块。文献[17]正是通过源语义块与目标语义块的

匹配来实现包装器的生成与维护。语义块的匹配是通过实体识别和聚类方法实现的。文献[17]还结合了 SG-WRAP 的模式指导的方法与 XPath-Wrapper 的 XPath 查找信息的技术来描述网页的领域特征，定义数据属性，达到精确抽取的目的。

3.2.3 对多种网页语言的支持

多数全自动包装器生成工具都是针对 HTML 网页信息的抽取。但是 HTML 的语法并不严格，因此 RoadRunner 将 HTML 转换为格式良好的 XHTML 文件后再进行处理。文献[17]为了能用 XPath 将网页中的数据抽取出来也将 HTML 转换为 XHTML。XPath-Wrapper 则增加了对 XML 文档的支持。由于网页标记语言的多样性，基于某种网页标记语言来定位目标数据存在较大的局限性，因而 ViDRE 设计了一种与网页语言无关的单纯依赖网页的可视化特征来定位目标数据的方法。

3.2.4 人工干预的程度

即使是全自动包装器生成技术也有少量的人工指导，特别是 SG-WRAP、XPath-Wrapper、ViDRE 和 WDE 在指定抽取模式、校正抽取规则或生成正、反训练样本的过程中都有不同程度的人工干预。而 SG-WRAP 专门设计人工干预的环节是为了达到精确抽取和按需抽取的目的。ViDRE 则提出“校正率”指标来评价人工干预的需求度。校正率是抽取误差与处理网页总数的比

率，校正率越高说明越需要人工修正抽取规则。

4 结语

通过对各包装器生成技术的分类、定性分析和特点概述，我们也许可以得出如下结论：

首先，全自动包装器生成技术的发展表现出抽取复杂数据的能力不断增强。从严格的结构化网页到松散的结构化网页，从单层结构数据到嵌套结构数据，从记录到字段，全自动包装器生成技术应对网页变化的能力和处理不同粒度信息的弹性不断增强；应用领域不断扩大，从单个领域到多主题领域再到不限制应用领域，其通用性和适应性不断增强。

其次，尽可能地降低用户使用上的负担，追求无人工指导的全自动包装器生成技术是各类包装器生成技术的目标。但是这并不排斥为了满足用户个性化定制需求而增加抽取模式的设定和抽取规则的校正等人机交互的接口。

第三，网页可视化特征处理技术的应用使得全自动包装器生成技术对网页语言的依赖大大降低。如何应对网页语言的不断变化以及如何加入更多网页可视化特征来定位目标数据，都将使这一技术成为今后的研究热点。

第四，不同于传统方法的模板依赖和参数假设，数据驱动的全自动包装器生成技术加强了对目标数据的语义描述使信息抽取变得更加有的放矢，但数据的属性描述受到了领域模式

局限性的约束，因而是今后需要进一步研究的技术。

最后，包装器的自动维护与其自动生成是同等重要的问题，但在以往文献中涉及较少，只在文献[17]中对网站改版前后的包装器维护问题与生成技术一并进行了考虑。面对网页的频繁变化，解决包装器的自动维护问题，增强包装器的弹性和适应性仍是今后需要解决的重点和难点。

本文的主要贡献在于初步构建全自动包装器生成技术的分类体系，并在此基础上对 2001 年以后出现的十五个全自动包装器生成技术进行了分类比较和定性分析，最后提出五点发展趋势。

参考文献

[1] Line Eikvil, Information Extraction from World Wide Web-A Survey[R]. Technical Report No.945, Norwegian Computing Center,1999.
[2] Alberto H. F., Altigran S., etc., A Brief Survey of Web Data Extraction Tools[J]. In:SIGMOD Rec.,2002,31(2):84-93.
[3] Crescenzi V., Mecca G., and Merialdo P.. RoadRunner:Towards Automatic Data Extraction from Large Web Sites[C].In:VLDB2001:109-118.
[4] Xiaofeng Meng, Hongjun Lu, etc.. SG-WRAP: A Schema-Guided Wrapper Generator Data Engineering[C].In:2002 Proceedings,18th International Conference on 2002: 331-332.
[5] Arasu A., Garcia-Molina H..Extracting Structured Data from Web Pages[C]. In:ACM SIGMOD Conference, 2003-06.
[6] Liu B．，Grossman R．，Zhai Y．Mining Data Records in Web Pages[C]. In:KDD2003:601-606.
[7] Wang J., Lochovsky F. H..Data Extraction and Label Assignment for Web Databases[C].In: Proceedings of the 12th international conference on World Wide Web ,2003:187-196.
[8] T. Anton. XPath-Wrapper Induction by Generalizing Tree Traversal Patterns[C]. In:LWA2005:126-133.
[9] Yanhong Zhai, Bing Liu. Automatic Wrapper Generation Using Tree Matching and Partial Tree Alignment[C]. In:2006 American Association for Artificial

Intelligence,2006.
[10] Yanhong Zhai , Bing Liu . Web Data Extraction Based on PartialTree Alignment[C]. In:WWW 2005, 2005-05.
[11] Kai Simon, Georg Lausen. ViPER:Augmenting Automatic Information Extraction with Visual Perceptions[C]. In:Proceedings of the 2005 ACM International Conference on Information and Knowledge Management (CIKM'05). Germany,2005-10:381-388.
[12] Hongkun Zhao, Weiyi Meng. Fully Automatic Wrapper Generation For Search Engines[C]. In:WWW 2005. Japan, 2005-05:66-75.
[13] Wei Liu, Xiaofeng Meng, Weiyi Meng. Vision-based Web Data Records Extraction[C]. In:Ninth International Workshop on the Web and Databases (WebDB2006), Chicago,2006-06:20-25.
[14] Justin Park，Denilson Barbosa. Adaptive Record Extraction From Web Pages[C].In:WWW2007,2007-05:1335-1336.
[15] 梅雪,程学旗,郭岩,张刚,丁国栋.一种全自动生成网页信息抽取 Wrapper 的方法[J].中文信息学报.2008，01.
[16] 杨少华，林海略，韩燕波.针对模板生成网页的一种数据自动抽取方法[J].软件学报.2008，19(2):209-223.
[17] 王仲远,艾静,孟小峰.一种数据驱动的 Wrapper 自动生成与维护方法[J].计算机研究与发展.2008，43.
[18] 李永丽,张玉良.一种基于后缀树的包装器自动生成方法研究[J].计算机工程与应用. 2007,43:114-118.
[19] Cai D., Yu S., Wen J., Ma W..Extracting Content Structure for Web Pages Based on Visual Representation[C]. In:APWeb, 2003: 406-417.

[作者简介]

李广建，Email：Ligj@pku.edu.cn
乔建忠，Email：qiaojianzhong@mail.las.ac.cn 或 j7z7q7@163.com