基于 Web 的信息抽取技术研究综述

蒲筱哥
（徐州师范大学，徐州 221116）

〔摘 要〕 本文在讨论 Web 信息抽取技术的发展历程、概念及其功能，Web 信息抽取技术方法的分类及技术特点分析，Web 信息抽取系统的构建研究及其性能评价的基础上，分析了当前研究存在的问题以及未来的研究方向。
〔关键词〕 网络信息资源；信息抽取；综述
〔**Abstract**〕 This article discusses the development course, the concept, the function, the classification and the technical characteristic analysis of Web information extraction, discusses the construction of Web information extraction system and the system evaluation, and analyzes the existence question of current research as well as the future research direction.
〔**Key words**〕 network information resource; information extraction; summarizes
〔中图分类号〕G250.73 〔文献标识码〕A 〔文章编号〕1008－0821（2007）10－0215－05

当前 Web 信息查询主要是通过各种搜索引擎进行，根据用户的查询请求，搜索引擎能找到相关的网页，但各网站的信息内容互相独立，搜索引擎的“网络爬虫”并不能收集到网上数据库内部的信息。据统计，目前 80％的网页属于隐藏网页（此类网页是由后台数据库生成），而搜索引擎无法从此类网页中获取数据。Web 信息抽取技术的研究就是在这样的背景下产生的。本文总结了 Web 信息抽取技术的研究现状，分析了当前研究的不足之处以及今后所需的研究。

1 Web 信息抽取技术的发展历程、概念及其功能

1.1 Web 信息抽取技术研究的发展历程

Web 信息抽取（Web Information Extraction：WIE）的前身是文本理解，最早开始于 20 世纪 60 年代中期，主要是从自然语言文本中获取结构化信息的研究，这被看作是信息抽取技术的初始研究。

从 20 世纪 80 年代末开始，信息抽取研究蓬勃开展起来，这主要有两个因素对其发展有重要的影响：一是在线和离线文本数量的几何级增加，另一个是“消息理解研讨会”（MUC，Message Understanding Conference）从 1987 年开始到 1998 年共举行了七届会议对该领域的关注和推动。MUC 由美国国防高级研究计划委员会（DARPA，the Defense Advanced Research Projects Agency）资助，其显著特点并不是会议本身，而在于对信息抽取系统的评测。MUC 系列会议对信息抽取这一研究方向的确立和发展起到了巨大的推动作用。MUC 定义的信息抽取任务的各种规范以及确立的评价体系已经成为信息抽取研究事实上的标准。近几年，信息抽取技术的研究与应用更为活跃。在研究方面，主要侧重于以下几方面：利用机器学习技术增强系统的可移植能力、探索深层理解技术、篇章分析技术、多语言文本处理能力、WEB 信息抽取（Wrapper）以及对时间信息的处理等等。在应用方面，信息抽取应用的领域更加广泛，除自成系统以外，还往往与其他文档处理技术结合建立功能强大的信息服务系统。至今，已经有不少以信息抽取技术产品为主的公司出现，比较著名的有：Cymfony 公司、Bhasha 公司、Linguamatics 公司、Revsolutions 公司等。

目前，除了强烈的应用需求外，正在推动信息抽取研究进一步发展的动力主要来自美国国家标准技术研究所（NIST）组织的自动内容抽取（ACE，Automatic Content Extraction）评测会议。这项评测从 1999 年 7 月开始酝酿，2000 年 12 月正式开始启动，从 2000 年到 2007 年已经举办过好几次评测。这项评测旨在开发自动内容抽取技术以支持对三种不同来源（普通文本、由自动语音识别 ASR 得到的文本、由光学字符识别 OCR 得到的文本）的语言文本的自动处理，研究的主要内容是自动抽取新闻语料中出现的实体、关系、事件等内容，即对新闻语料中实体、关系、事件的识别与描述。与 MUC 相比，目前的 ACE 评测不针对某个具体的领域或场景，采用基于漏报（标准答案中有而系统输出中没有）和误报（标准答案中没有而系统输出中有）为基础的一套评价体系，还对系统跨文档处理（Cross－document processing）能力进行评测。这一新的评测会议将把信息抽取技术研究引向新的高度。

国内对中文信息提取系统的研究起步较晚，还集中在命名实体识别方面，遵照 MUC 规范的完整的中文信息提取系统目前还处于探索阶段。Intel 中国研究中心在 ACL－2000 上演示了他们开发的一个抽取中文命名实体以及实体间关系的系统。在 MUC－6 和 MUC－7 上，增加了中文系统的评测项目，国立台湾大学（National Taiwan University）和新加坡肯特岗数字实验室参加了 MUC－7 中文命名实体识别任务的评测，测试了中文命名实体（人名、地名、时间、事件等名词性短语）的识别，取得了与英文命名实体识别系统相近的性能。当然这只是对中文信息提取作了比较初步的工作，并不能真正进行中文信息提取。另外，北

收稿日期：2007－08－29
作者简介：蒲筱哥（1972－），男，中山大学资讯管理系研究生毕业，徐州师范大学图书馆馆员，研究方向：网络信息资源开发与利用。

京大学计算语言所对中文信息提取也作了比较早的和比较系统的探讨，承担了两个有关中文信息提取项目的工作，即自然科学基金项目“中文信息提取技术研究”和IBM——北大创新研究院项目“中文信息提取系统的设计与开发”。其目标是研究中文信息提取中的一些基础性和关键性的问题，为开发实用的信息提取技术提供理论指导，并具体探讨信息提取系统设计的各个环节。

1.2 Web信息抽取的概念及其功能

信息抽取的概念有多种描述方式，1997年Proteus工程的创建者Grishman描述信息抽取的概念：“信息抽取涉及到为从文本中选择出的信息创建一个结构化的表示形式（比如：数据库）”，微软亚洲研究院2005年信息抽取技术暑期研讨班将信息抽取的概念描述为：“信息抽取是抽取和链接基于用户详细说明的相关信息的过程”。结合种种对信息抽取概念的描述，以及过去20年里一系列的消息理解会议（Message Understanding Conference，MUC）对信息抽取技术的讨论，综观各定义，可以将Web信息抽取的概念界定为：Web信息抽取（Web Information Extraction：WIE）就是从网页文本中抽取指定的一类信息（事件、事实）并将其形成结构化的数据填入一个数据库中供用户查询使用的过程。

Web信息抽取技术的核心是能够从Web页所包含的无结构或半结构的信息中识别用户感兴趣的数据，并将其转化为更为结构化、语意更为清晰的格式。输入信息抽取系统的是原始文本，输出的是固定格式的信息点。信息点从各种各样的文档中被抽取出来，然后以统一的形式集成在一起。这就是Web信息抽取的主要功能。IE系统中的关键组成部分是一系列的抽取规则或模式，其作用是确定需要抽取的信息。

Web信息抽取的内容一般可以分为这样几个方面：命名实体的抽取、与模板有关的内容信息抽取、各个实体之间关系的抽取和预置事件的信息抽取。

命名实体的抽取：它包括组织机构、人名、地名的抽取，时间、日期、钱币和百分数的抽取、专有名词的抽取、隐含指代名词和集合名词的抽取。命名实体的自动抽取能力已近似于人工抽取：查准率达到了70%以上，查全率是60%。

模板内容信息的抽取：用户预先设置模板，自动抽取用户关心的详细内容，反映时间、地点、人物和发生的事件。

实体关系信息的抽取：比如某些疾病的因果关系。

预置事件信息的抽取：比如公司宣布破产、合并的消息、原因等等。事件信息抽取的查准率目前维持在50%～60%。

2 Web信息抽取技术方法的分类及技术分析

2.1 根据Web信息源划分

Web信息源可以分为3类，即自由文本、结构化文本、半结构化文本，但以半结构化文本为主。针对这三种文本，信息抽取也分为三种类型：（1）从自由格式的文本中抽取出所需要的信息内容，自由文本的抽取技术可分为三类：基于NLP（自然语言处理）的方式、基于规则的方式和基于统计学习的方式。基于NLP的方式是早期的信息抽取方法，一般效率较低，现已较少使用。基于规则的方式是一种知识工程的方法。在早期，一般以手工的方式设置抽取规则。随着应用范围的扩大，手工获取规则成为知识工程的瓶颈。近期大量语料库的涌现，为规则的自动学习和获取提供了可能，这使得机器学习的方法在规则的（半）自动获取中得到广泛应用，基于规则的方式成为当前信息抽取的主流。基于统计学习的方式主要有基于HMM（隐马尔可夫模型）的方法等，由于HMM的参数可通过训练获得，这种方式的可移植性较好；（2）从半结构化的文本中，抽取出所需要的信息内容，对于半结构化文本传统的采用自然语言处理技术的信息抽取系统已经不适用了，其抽取模式经常是基于标记和分界符。如HTML标记等，句法和语义信息只是在一定范围内被使用；（3）从结构化的文本中抽取出所需要的信息内容，这种信息抽取任务最为简单，结构化文本是指数据库中文本信息或遵循预先定义的而且严格的格式的文本，这样的信息易于使用格式描述进行抽取，对这种事先格式已知的文本进行信息抽取通常用比较简单的技术就可以了。

2.2 根据包装器（Wrapper）不同原理可以分为以下两类：

2.2.1 基于层次结构的信息抽取归纳方法

如WHIRL、Ariadne、CiteSeer等是基于层次结构的Wrapper归纳方法。基于层次结构的Wrapper归纳方法引入嵌套目录描述方法（Embedded Catalog，简称EC），该方法将页面内容按照层次结构树（EC树）的形式加以描述。EC树的叶节点用以描述用户感兴趣的相关数据；EC树的内部节点用以描述由多个项目组成的列表，其中每个项目既可以是叶节点，也可以是项目列表（项目嵌套）。EC树中每一条边均与一个抽取规则相关联，每一个列表节点与一个列表循环规则相关联，根据从根节点到相应叶节点的路径，依次从父节点抽取路径上的每个子节点，Wrapper就能够从页面中抽取任何用户感兴趣的项目。

2.2.2 基于概念模型的多记录信息抽取方法

即对特定WWW数据源研制相应的Wrapper。通过记录识别获得记录相对应的信息块格式，利用Wrapper进行有效的记录抽取。步骤是设计构造描述特定内容的本体模型（Ontology Model），并由此产生一个数据库模式以及产生有关常量/关键字的匹配规则。抽取信息时系统调用记录抽取器将页面分解为若干单个记录信息块，并除去其中的标记，利用由分析器（Parser）产生的匹配规则，从所获得的单个信息块中抽取有关的对象以及它们之间的关系，并将他们存入数据记录表（Data－Record Table）；最后利用启发知识，并根据有关的数据库模式，将所获得的数据填入相应的数据库中。利用启发知识，在所抽取的常量与关键字之间建立关联，并利用层次结构中的约束，来决定如何构造数据库中的有关记录内容。

2.3 根据自动化程度就可以分为人工方式的信息抽取、半自动方式的信息抽取和全自动方式的信息抽取3大类。

2.4 根据各种信息抽取工具所采用的原理将现有的工具分为5类：基于自然语言处理（NLP）方式的信息抽取、包装器（Wrapper）归纳方式的信息抽取、基于ontology方式的信息抽取、基于HTML结构的信息抽取和基于Web查询的信息抽取。

2.4.1 基于自然语言处理（Natural Language Processing，

NLP）方式的信息抽取自然语言的处理过程一般可归为：语音、词、词形、语法、语义、篇章、语用 7 个不同的抽象级别。这类信息抽取主要适用于源文档中包含大量文本的情况（特别针对于合乎文法的文本），在一定程度上借鉴了自然语言处理技术，利用子句结构、短语和子句间的关系建立基于语法和语义的抽取规则实现信息抽取。NLP 方式难点在于：信息抽取速度太慢，信息抽取与文本理解之间存在较大的差别——信息抽取只关心相关的内容，而文本理解则要能体会作者的细微用意和目的。目前采用这种原理的典型的系统有 RAPIER，SRV，WNISK。

2.4.2 包装器（Wrapper）归纳方式的信息抽取

包装器归纳方式的信息抽取是根据事先由用户标记的样本实例应用机器学习方式的归纳算法，生成基于定界符的抽取规则。其中定界符实质上是对感兴趣语义项上下文的描述，即根据语义项的左右边界来定位语义项。该类信息抽取方式和基于自然语言理解方式的信息抽取技术最大的不同在于仅仅使用语义项的上下文来定位信息并没有使用语言的语法约束。采用这种原理的典型的系统有 STALKER，SOHTMEALY，WIEN。

2.4.3 基于 ontology 方式的信息抽取

该类信息抽取主要是利用对数据本身的描述信息实现抽取，对网页结构的依赖较少。由 Brigham Yong University 信息抽取小组开发的信息抽取工具中采用了这种方式，另外 QU IXOTE 也采用了这种方式。

2.4.4 基于 HTML 结构的信息抽取

该类信息抽取技术的特点是，根据 Web 页面的结构定位信息。在信息抽取之前通过解析器将 Web 文档解析成语法树，通过自动或半自动的方式产生抽取规则，将信息抽取转化为对语法树的操作实现信息抽取。采用该类技术的典型系统有 LIXTO 等。

2.4.5 基于 Web 查询的信息抽取

使用 Web 的相关技术解决 Web 的问题称为 Web 技术风范。上述的信息抽取工具，采用了不同的原理，抽取规则的形式和感兴趣信息的定位方式也各不相同，因此均不具有通用性，具有 Web 技术风范的信息抽取，将 Web 信息抽取转化为使用标准的 Web 查询语言对 Web 文档的查询，具有通用性。采用该类技术的典型的系统有：Web－OQL 以及自主开发的原型系统 PQAgent。

3 Web 信息抽取系统的构建研究及其性能评价

3.1 Web 信息抽取系统的构建研究

3.1.1 Web 信息抽取系统的体系结构

Hobbs 曾提出个信息抽取系统的通用体系结构，他将信息抽取系统抽象为“级联的转换器或模块集合，利用手工编制或自动获得的规则在每步过滤掉不相关的信息，增加新的结构信息”。

Hobbs 认为典型的信息抽取系统应当由依次相连的十个模块组成：

（1）文本分块：将输入文本分割为不同的部分——块。

（2）预处理：将得到的文本块转换为句子序列，每个句子由词汇项（词或特定类型知语）及相关的属（如词类）组成。

（3）过滤：过滤掉不相关的句子。

（4）预分析：在词汇项（lexical Items）序列中识别确定的小型结构，如名词知语、动词知语、并列结构等。

（5）分析：通过分析小型结构和词汇项的序列建立描述句子结构的完整分析树或分析树片段集合。

（6）片段组合：如果上一步没有得到完整的分析树，则需要将分析树片段集合或逻辑形式片段组合成整句的一棵分析树或其他逻辑表示形式。

（7）语义解释：从分析树或分析树片段集合生成语义结构、意义表小或其他逻辑形式。

（8）词汇消歧：消解上一模块中存在的歧义得到惟一的语义结构表示。

（9）共指消解或篇章处理：通过确定同一实体在文本不同部分中的不同描述将当前句的语义结构表示合并到先前的处理结果中。

（10）模板生成：由文本的语义结构表示生成最终的模板。

当然，并不是所有的信息抽取系统都明确包含所有这些模块，并且也未必完全遵循以上的处理顺序，比如（6）、（7）两个模块执行顺序可能就相反。但每个信息抽取系统应当包含以上模块中描述的功能。

3.1.2 Web 信息抽取系统的工作过程

信息抽取系统是利用一种由事件名称（Event）、日期（date）、时间（time）、地点（location）等槽（slot）组成的信息模式，对报道中相应的内容进行匹配，并正确填满各槽的内容。一般而言，一个典型的信息抽取系统的工作过程主要包括如下几个步骤：

（1）用一组信息模式描述感兴趣的信息。系统可以针对某一领域的信息特征预定义好一系列的信息模式，存放在模式库中供用户选用。

（2）对文本进行“适度的”词法、句法及语义分析，通常包括识别特定的名词短语（人名、机构名、产品名、事件、地点等）和动词短语（事件描述、事实陈述）。这需要使用合适的词典、构词规则库等知识库的支持。

（3）使用模式匹配方法识别指定的信息（即找出信息模式的各个部分）。

（4）进行上下文关联、指代、引用等分析和推理，确定信息的最终形式。

（5）输出结果（例如生成一个关系数据库或给出自然语句陈述等）。

出于效率的考虑，有的信息提取系统还包括一个预处理过程，目的在于过滤掉与提取目标不相关的文本。

3.1.3 Web 信息抽取系统实现的方法

IE 系统设计主要有两大方法：一是知识工程方法（Knowledge Engineering Approach），二是自动训练方法（Automatic Training Approach）。

知识工程方法主要靠手工编制规则使系统能处理特定知识领域的信息抽取问题。这种方法要求编制规则的知识工程师对该知识领域有深入的了解。这样的人才有时找不到，且开发的过程可能非常耗时耗力。

自动训练方法不一定需要如此专业的知识工程师。系统主要通过学习已经标记好的语料库获取规则，任何对该知识领域比较熟悉的人都可以根据事先约定的规范标记语

料库，经足够数量的数据实验后的系统能处理没有见过的新文本。这种方法要比知识工程方法快，但需要足够数量的实验数据，才能保证其处理质量。

3.2 Web信息抽取系统性能评价

Web信息抽取技术的评测起先采用经典的信息检索（IR）评价指标，即回召率（Recall）和查准率（Precision），但稍稍改变了其定义。经修订后的评价指标可以反映IE可能产生的过度概括现象（Over－generation），即数据在输入中不存在，但却可能被系统错误地产生出来（Produced）。就IE而言，回召率（recall）可粗略地被看成是测量被正确抽取的信息的比例（fraction），而抽准率（precision）用来测量抽出的信息中有多少是正确的。计算公式如下：

P＝抽出的正确信息点数/所有抽出的信息点数

R＝抽出的正确信息点数/所有正确的信息点数

两者取值在0和1之间，通常存在反比的关系，即P增大会导致R减小，反之亦然。评价一个系统时，应同时考虑P和R，但同时要比较两个数值，毕竟不能做到一目了然。当比较两个不同信息抽取系统的性能时，一般使用这两个指标的综合值F度量。

$$F=[(B2+1)\times P\times R]/[(B2\times P)+R]$$

其中：P为精度，R为召回率，B为对精度的偏重量，其中B是一个预设值，决定对P侧重还是对R侧重，通常设定为1，这样用F一个数值就可看出系统的好坏。

4 当前研究存在的问题及今后的研究

4.1 当前研究存在的问题

Web信息抽取技术目前已基本成熟，但知识的自动获取实际上仍没有达到完全自动，大部分信息抽取系统只是把原先由领域专家完成的任务转化为用户的任务。在构建通用的知识学习器方面，进行了有益的探讨，但效果不是很理想，当前基于Web的IE系统只能处理特定类型的文本和只能获得部分的精确度，仍面临很多问题。

4.1.1 当前影响Web信息抽取技术广泛应用的两个最主要的因素是：系统性能和系统可移植能力，如何解决好这两方面的问题将决定Web信息抽取系统的发展水平，人工智能研究者一直致力于建造能把握整篇文档的精确内容的系统。这些系统通常只在很窄的知识领域范围内运行良好，向其他新领域移植的性能却很差。

4.1.2 Web信息抽取系统的抽取效率和抽取的准确性有待进一步提高。

4.1.3 目前英文系统在命名实体和实体关系识别方面已达到或接近实用的水平。但在真正的信息提取方面则还有许多问题需要探索。可以看到这些问题中的大部分都涉及到了自然语言处理中的核心难题。

4.1.4 定义包含从文本中抽取的重要信息的模版是一个十分困难和复杂的问题，特定流派的文本（如医学结论、科学论文、政策报告等）具有特定的词汇、语法和篇章结构。系统分词与词性标注过程中存在歧义问题，语义特征标注、篇章句法分析等也是一个需要进一步研究的课题。

4.1.5 系统在适应不同子语言特征、不同类别的文本方面有待提高。系统应能处理特定语言结构和多语种文本，基于Web的文档可能与新闻报纸之类的文本有着强烈的差别，必须能适应不同的情况。

4.1.6 与国外的Web信息抽取系统相比，中文信息抽取系统的研究还有很大差距。

4.2 今后的研究

针对当前研究存在的问题，今后如何将提高Web信息抽取系统抽取范围的全面性；如何简化学习过程，提高自动化程度；如何提高系统对新网页的适应性，增强系统对Web信息抽取的适应性；如何加强对已有抽取规则的归纳，提高系统的抽取效率和准确性；Web上的信息和网页结构处于不断的更新和变化中，因此应如何感知Web信息和结构的更新变化；目前的Web信息抽取工具一般都是通过学习之后可以对结构相似的一类网页进行抽取，因此应如何判断结构相似；如何提高系统的性能、可移植性的设计以及适应多语种的能力；在中文Web信息抽取系统的研究方面，应如何借鉴国外比较成熟的系统构建技术，并结合汉语的特殊性，充分利用一些基础的汉语研究成果来构建高效、精确的中文Web信息抽取系统；这些问题都是今后Web信息抽取技术研究的热点问题。

5 结 语

Web信息抽取技术是一个年轻的研究领域，尽管目前该领域研究已经取得了一定的进展，但仍然存在一些问题。在一个新领域上建立信息抽取系统还需要许多该领域的专家和熟悉NLP系统的计算语言学家的共同努力。随着计算机网络在国内的迅猛发展，Web信息抽取技术会变得越来越重要，希望有更多更好的技术能够应用到该领域，从而使处理动态的海量信息的Web信息抽取技术的自动化程度及精准度越来越高。

参 考 文 献

[1] Lawrence S, Giles C L. Searching the world wide web [J]. Science, 1998, 280 (4): 98－100.

[2] Grishman R, Sundheim B. message Understanding Conference on Computational Linguistics COLING－96, 1996－08.

[3] http://www.cymfony.com/index.html [EB]. 2007.5.

[4] http://www.bhasha.com/[EB]. 2007.5.

[5] http://www.linguamatics.com/index.html [EB]. 2007.5.

[6] http://www.revsolutions.com/index.html [EB]. 2007.5.

[7] http://www.itl.nist.gov/iad/894.01/tests/ace [EB]. 2007.5.

[8] 邓尚民，孙玉伟. 信息抽取系统的研究现状 [J]. 现代图书情报技术，2006, (3): 55－58, 81.

[9] Ralph Grishman. Information extraction: Techniques and Challenges. InMaria Teresa Pazienza, editor, Information Extraction. Sp ringer－Verlag, Lecture Nots in Artificial Intelligence, Room, 1997. 61－67.

[10] Ralph Grishman and Beth Sundheim. Message Understanding Conference－6: A Brief History. In Proceedings of 16 th International Computational Linguistics. 1996. 31－35.

[11] Chinchor N. overview of MUC－7/MET－2. in: Proceedings of the SeventhMessage Understanding Conference, 1998. 23－28.

[12] Proceedings of the Third Message Understanding Conference (MUC－3). Morgan Kaufmann, May, 1991.

[13] Proceedings of the Fourth Message Understanding Conference (MUC－4). Morgan Kaufmann, 1992.

[14] Proceedings of the Fifth Message Understanding Conference (MUC－5). Baltimore, MD, August, 1993. Morgan Kaufmann.
[15] Proceedings of the Sixth Message Understanding Conference (MUC－6). Columbia, MD, November, 1995. Morgan Kaufmann.
[16] S. Soderland. Learning Information Extraction Rules for Semistructured and Free Text. Machine Learning, 1999.
[17] http://www.Cs.cmu.edu/~ref/mlim/chapter3.html [EB/OL]. 2007. 3.
[18] Christopher S G Khoo, Syin Chan, Yun Niu1 Extracting Causal Knowledge from a Medical Database Using Graphical Patterns [C]. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics HongKong: 128 Oct. 2000 336－343.
[19] http://www.Cs.cmu.edu/~ref/mlim/chapter3.html [EB/OL]. 2007. 3.
[20] 李向阳, 苗壮. 自由文本信息抽取技术 [J]. 情报科学, 2004, (7): 815－819.
[21] 吴振慧. Web 信息抽取的研究 [J]. 电脑知识与技术, 2005, (3): 21, 24.
[22] LAENDER A, RIBEIRO－NETO B, SILVA A. A brief survey of web data extraction Tools [J]. SIGMOD Record, 2002, 31 (2): 84－93.
[23] CALIFF M, MOONEY R. Relational Learning of pattern－match rules for information extraction [Z]. In Proceedings of the Sixteenth National Conference on Artificial Intelligence and Eleventh Conference on Innovative Applications of Artificial Intelligence, Orlando, Florida, 1999.
[24] FREITAC D. Machine learning for information extraction in informal domains [J]. Machine Learning, 2000, 39 (2/3): 169－202.
[25] SODERLAND S. Learning information extraction rules for semi－structured and Free Text [J]. Machine Learning, 1999, 34 (1－3): 233－272.
[26] MUSLEA I, MINTON S. KNOLCX－K C. Hierarchical wrapper induction for semistructured information sources [J]. Autonomous Agents and Multi－Agent Systems, 2001, 4 (1/2): 93－14.
[27] CRAIG A, KNOBLOCK, KRISTINA L, et al. Accurately and reliably extracting data from the web: A machine learning approach [J]. Data Engineering Bulletin, 2000, 23 (4): 33－1.
[28] HSUC N, DUNG M. (Generating finite－state transducers for semi－structured data extraction from the Web [J]. Information System, 1998, 23 (8): 521－538.
[29] KUSHMF: KICK N. Wrapper induction: efficiency and expressiveness [J]. Artificial Intelligence Journal, 2000, 118 (1/2): 15－68.
[30] EMBLEY D, CAMPBELL D, JIANG S, et al. Conceptual－model－based data extraction from ultiple record web pages [J]. Data and Knowledge Engineering, 1999, 31 (3): 227－251.
[31] CHRISTINA YIP CHUNG, MICHAEL GERTZ, NEEL SUNDARESAN. Reverse engineering for Web data: From visual to semantic structures [Z]. In Proceedings of 18th International Conference on Data Engineering, San Jose, California, 2002.
[32] CHRISTINA YIP CHUNG, NEEI－SUNDARESAN. Quixote: Building XML repositories from topic specific web documents [Z]. In Fourth Int. Workshop on the Web and Databases, 2001.
[33] ROBERT BAUMGART`NER, SERGIO FLESCA, GEORG GOTTLOB. Supervised wrapper generation with lixto [Z]. Proceedings of 27th International Conference on Very Large Database, Roma, Italy, 2001. 22－26.
[34] 陈少飞. Web 信息抽取技术研究进展 [J]. 河北大学学报, 2003, (1): 106－112.
[35] Hobbs J. The generic Information Extraction System [C]. In: Proceeding of the Fifth Message Understanding Conference MUC－5, Morgan Kaufman, 1993: 87－91.
[36] 李保利, 等. 信息抽取研究综述 [J]. 计算机工程与应用, 2003, (10): 1－5.
[37] Cardie C. Empirical methods in information extraction. AI Magazine, 1997, 18 (4): 65－79.
[38] 李晶, 陈恩红. Web 信息抽取 [J]. 计算机科学, 2003, (6): 78－81.
[39] Line Eikvil, 陈鸿标. 网上信息抽取技术纵览 [EB]. 2003, www.fullsearcher.com/down/InformationExtraction/1.doc, 2007. 5.
[40] R Gaizauskas, Y. Wilks. Information Extraction: Beyond Document Retrieval. Computational Linguistics and Chinese Language Processing, 1998, 3 (2): 17－60.
[41] K. Zechner. A Literature Survey on Information Extraction and Text Summarization. Term paper: CarnegieMellonUniversity, 1997.

(上接第 124 页)
部门管理“其馆藏设施、基础硬件、经费来源、人员配置等都不尽相同, 甚至有较大差异。这样一来, 信息资源输出与接收的矛盾就产生了, 肯定出现因基础配置不同而带来的差异。规模大、基础好、设施现代化的图书馆在信息资源共享流程中面临的问题是: 信息资源输出大于接受; 相反, 规模小、基础差、设施落后的图书馆则是: 信息资源接受大于输出。与之带来的诸多劳务等成本费用因此出现悬殊, 在缺乏利益平衡机制的状态下, 投入少收益大者自不言表, 而投入大收益少的因得不到相应回报, 必然伤害到他们参与信息资源共享的积极性, 久而久之态度难免消极”。

综上所述, 本文笔者只是对部分中小型馆的文献资源建设作了一些调查、统计、分析, 意图起到抛砖引玉的作用。对提高中小型馆文献资源建设对应策略, 由于篇幅关系, 在另一篇文章进行探讨研究。

参 考 文 献

[1] 卢勇. 数字时代中小型图书馆文献资源建设 [J]. 图书馆研究与工作, 2005, (3): 36.
[2] 廖柯夫. 社区资源共享工程构建 [J]. 四川图书馆学报, 2006, (2): 18.