一种基于网页信息抽取的OA期刊资源采集方法研究

黄政，张学福
（中国农业科学院农业信息研究所，北京 100081）

摘要：本文结合开放获取期刊（Open Access Journal，OA期刊）资源特点，针对无法通过OAI-PMH协议进行资源采集的OA期刊，提出一种基于网页信息抽取的资源采集策略。本文从网页资源描述的角度总结OA期刊资源特点并对其分类。基于网页信息抽取方法在OA期刊资源采集适用性，提出一种基于OA期刊网页元数据抽取的采集方法，并在此方法的基础上设计了采集系统。通过对国内外不遵循OAI-PMH协议的10本OA期刊的网站实证采集，得到45 785篇论文的元数据，证明该采集方法能有效地应用于此类资源采集。研究丰富了OA期刊资源采集方式，对不遵循OAI-PMH协议的OA期刊资源采集提供方法借鉴。

关键词：OA期刊；OA期刊资源采集；网页信息采集；OA期刊资源采集系统
中图分类号：G250
DOI：10.3772/j.issn.1673-2286.2017.05.004

开放获取期刊（Open Access Journal，OA期刊）是经过同行评审，且在网络上可免费获取的期刊。OA期刊资源主要包括期刊元数据、论文元数据以及论文全文等。该类资源分布广泛，且经过同行评审，具有重要的学术价值。目前，OA期刊资源采集方法主要有两种：一种是针对遵循OAI-PMH协议的OA期刊，采用OAI-PMH协议的方法对资源进行采集，该方法在此类资源采集应用中较成熟；另一种是对于部分不遵循OAI-PMH协议的OA期刊，通常采用网页信息抽取方法。然而，由于OA期刊资源在网页中存在组织形式不一、揭示粒度多变，且网页结构变化多样等特点，这为此类期刊资源采集带来了一定挑战。本文将从OA期刊资源特点出发，对网页信息采集方法和采集工具在OA期刊资源采集中的适用性进行对比分析，针对无法通过OAI-PMH协议进行资源采集的OA期刊，提出一种基于网页信息抽取的资源采集策略。以期既能丰富OA期刊资源采集方式，也能对不遵循OAI-PMH协议的OA期刊资源采集提供指导，提高资源采集效率。

1 文献回顾

OA期刊资源采集的研究现状可以从网页信息采集、开放获取资源采集和OA期刊资源采集三个角度进行分析。

在网页信息采集方面，根据采集包装器形成方式将采集方法分为：（1）基于自然语言处理的网页信息抽取，即将网页信息作为文本，使用自然语言处理技术来抽取网页信息；（2）基于本体的网页信息抽取，即将网页正文信息与构建的本体集进行比较并计算相关度，从中抽取相关度高的信息；（3）基于包装器归纳方式的网页信息抽取，即对有标注的样本网页采用机器学习算法来归纳抽取规则，并利用该规则抽取其他网页信息；（4）基于HTML页面结构分析的网页信息抽取，即将网页解析为结构树，对比多个网页，进而构建抽取信息的正则表达式采集网页中的信息；（5）基于Web查询的网页信息抽取，即先将网页进行解析，再使用类似数据库查询语句对网页信息进行采集[1-4]。

在开放获取资源采集方面，有学者对不同类型的开放获取资源采集进行了研究。朱江等研究开放会议资源采集，利用用户推荐和人工收集方式对Web环境下的开放会议资源进行采集，采用文本识别的方式抽取非结构化文本格式的会议文献开放资源[5]；王思丽等根据开放知识资源的不同数据来源提出不同的自动采集策略，包括基于OAI-PMH协议的元数据采集策略、

基于抽取动态网页的元数据采集策略和基于解析RSS源接口的元数据采集策略[6]。对开放获取资源采集方法的研究也越来越全面和深入，对所采集资源从一概而论变为分类制定采集策略，开放资源采集方法研究逐步从人工采集过渡到自动采集。除方法层面的研究外，有学者也从系统层面展开研究。宋辰对科技情报采集系统进行研究，指出当前科技情报采集工具难以满足情报资源采集需求的原因之一在于收费系统需要花费大量财力和人力，并且系统使用和维护困难[7]。

在OA期刊资源采集方面，基于OAI-PMH协议的元数据采集方法对于主要局限于遵循OAI-PMH协议的OA期刊，资源采集的应用已十分成熟[8-12]。针对OA期刊网页中展示的资源主要是先通过人工分析网页结构，再使用网页解析工具来对资源进行采集[13]，该方法主要以人工考察分析网页结构为主，需要采集者具有一定的计算机专业背景，而且工作量大，不适合对大量期刊资源采集。OA期刊资源属于网络资源的一种，对不遵循OAI-PMH协议的OA期刊，可以借鉴网页信息采集方法。文本将从网页信息采集的角度出发，结合OA期刊资源特点，对不遵循OAI-PMH协议的OA期刊资源采集策略进行研究，以满足此类OA期刊资源采集需求。

2 不遵循OAI-PMH协议的OA期刊资源采集方法研究

2.1 OA期刊资源的特点与分类

OA期刊分为遵循OAI-PMH协议和不遵循OAI-PMH协议两种，但所有的OA期刊都是通过网页对资源进行描述和展示，且描述和展示的方式差异较小，故本文分析的OA期刊资源特点适用于所有类型。

2.1.1 OA期刊资源的特点

（1）描述粒度细。OA期刊资源的元数据包含众多字段，如文章标题、中英文关键词、中英文摘要、作者、机构、期刊名、年、卷、期等。相比于其他网络资源，OA期刊资源元数据描述粒度更细。

（2）展现形式多样。OA期刊资源的元数据字段众多，而这些字段通常是以不同的组织形式展现在网页中。部分元数据字段在网页中是按照单个字段进行展示，如文章标题、摘要等；而部分元数据是多个字段组合成一条文本信息进行展示，如文章的年、卷、期。

（3）描述载体结构多变。在对国内OA期刊资源调研过程中发现，部分OA期刊网站的资源展示页面，在不同时期采用不同的网页模板。在结构发生变化的开放获取资源网站中，一般会存在1—3套不等的网页模板；而其他网络资源，如电商平台、论坛等通常采用统一的网页模板。

2.1.2 OA期刊资源分类

OA期刊资源以不同的组织形式在不同网页中进行展示，本文根据OA期刊资源在网页中的组织形式，将其分为单一型资源和组合型资源。

单一型资源指网页中一个HTML标签仅展示一个元数据字段信息的资源，如期刊名称、文章标题、摘要、关键词、全文获取链接等。此类资源信息揭示简单明了、层次清晰。

组合型资源指网页中一个HTML标签封装多个期刊元数据字段信息的资源，多个字段通常是组合成一个文本信息进行展示，如期刊的年、卷、期字段等。组合型资源的文本信息由固定字段按照一定的形式组合而成，具有一定的结构性，为半结构化文本。

2.2 现有网页信息采集方法的特点及适用性分析

2.2.1 现有网页信息采集方法特点分析

现有网页信息采集方法主要分为基于自然语言处理的网页信息抽取、基于本体的网页信息抽取、基于包装器归纳方式的网页信息抽取、基于HTML页面结构分析的网页信息抽取以及基于Web查询的网页信息抽取。5种采集方法特点对比分析结果如表1所示。

由表1可见，5种网页信息采集方法采用不同方式来保证资源采集的准确性。如基于包装器归纳方式的网页信息抽取方法需要对样本进行标注，通过机器学习归纳抽取规则来提高采集准确率；基于Web查询的网页信息抽取方法通过对网页分析，编写合适查询语句来准确定位页面中资源。不同的Web信息采集方法由于采集方式不同，适用于不同类型的网页资源采集。如基于自然语言处理的网页信息抽取方法适用于大量

表 1 5种网页信息采集方法特点对比分析

方法名称	采集方式	优 点	缺 点
基于自然语言处理的网页信息抽取方法	依赖自然语言处理技术和语意字典，建立短语间关系，所产生的抽取规则通常基于语法约束和语义约束	适用于大量文本信息且句子完整的网页	HTML页面内容通常不包含完整句子，抽取规则难以形成，处理速度慢
基于本体的网页信息抽取方法	建立本体集，对网页进行解析，将处理后的网页信息传递给计算模块，根据本体集来计算信息的相关度，进而保留相关信息	不依赖网页结构，适用于特定主题或领域的元数据	依赖专家建立强大的本体集，且工作量大
基于包装器归纳方式的网页信息抽取方法	可手动或自动形成抽取规则。手动形成需对数据源十分了解，自定义规则；自动形成需标记样本，通过机器学习训练包装器。通常一种信息源使用一个包装器处理	手动形成抽取规则，抽取准确度高；自动形成抽取规则速度快、易于维护。适用于结构清晰的网页，可以对较复杂的元数据进行采集	手动形成抽取规则耗时多，自动形成抽取规则需对网页结构非常了解
基于HTML页面结构分析的网页信息抽取方法	将网页解析成结构树，通过自动或半自动方式产生抽取规则，定位主题信息，将网页信息抽取转化成对解析树的操作	无须用户标注，适用于结构清晰且信息单一的元数据	准确率低，且需要大量样本网页进行训练
基于Web查询的网页信息抽取方法	将Web文档解析成一棵抽象的HTML语法树，根据信息在页面代码中的唯一标记，写出合适的类似于SQL语句的Web查询语言对语法树进行查询	通用性强，准确率高，不需要样本训练，适用于信息单一且标识明确的元数据	依赖网页结构，需要对网页进行分析，寻找合适的查询语句

文本信息抽取，基于本体的网页信息抽取方法适用于特定领域的信息抽取。

2.2.2 网页信息采集方法对OA期刊资源采集的适用性分析

与传统网页信息采集不同的是，OA期刊资源采集更注重网页内部元数据的过滤和抽取，网页元素采集准确率是衡量采集方法适用性的基本指标。每本OA期刊的网页结构各不相同，因此采集方法需要具有很好的灵活性，以应对不同网页结构的OA期刊资源采集。单一型资源采集类似于普通网页元数据采集，仅抽取网页标签对封装的信息；而组合型资源除抽取网页标签对封装的文本信息外，还需要对文本信息进一步采集，抽取文本信息中的单个资源信息。因此，文本信息抽取是采集OA期刊资源组合型元数据资源的主要方式。综合而言，采集准确率和方法灵活性是衡量方法适用性的基础，而文本信息处理是全面采集OA期刊资源的衡量指标。通过对5种网页信息采集方法特点以及优缺点分析，结合5种方法在OA期刊资源采集上的应用，对5种方法适用性对比分析如表2所示。

表 2 5种网页信息采集方法适用性对比分析

采集方法	衡量指标		
	采集准确率	方法灵活性	文本信息处理
基于自然语言处理的网页信息抽取方法	否	否	是
基于本体的网页信息抽取方法	是	否	是
基于包装器归纳方式的网页信息抽取方法	是	否	是
基于HTML页面结构分析的网页信息抽取方法	是	否	否
基于Web查询的网页信息抽取方法	是	是	否

通过对5种网页信息采集方法的适用性分析，得出两个结论。（1）现有主要的网页信息采集方法无法单独完成OA期刊资源采集工作。基于Web查询的网页信息抽取方法具备采集准确率和方法灵活性特征，但无法对文本信息进行处理。而其他4种方法无法兼备采集

准确率和方法灵活性。在文本信息处理方面，虽然基于本体的网页信息抽取方法和基于包装器归纳方式的网页信息采集方法通过构造本体集或构造包装器能够对文本中的信息抽取，但基于自然语言处理的信息抽取方法能更灵活、准确地抽取文本信息。（2）OA期刊资源采集方法需要综合网页信息采集方法的功能。虽然现有网页信息采集方法无法完成OA期刊资源的完整性采集，但基于Web查询的网页信息抽取方法和基于自然语言处理的网页信息抽取方法分别具备OA期刊资源采集的基础性指标和全面性指标，OA期刊资源采集方法需要综合这两种Web信息采集方法的功能，实现OA期刊资源灵活、准确和全面地采集。

2.3 基于网页信息抽取的OA期刊资源采集方法

通过分析5种网页信息采集方法的特征，以及各方法在OA期刊资源采集的适用性，认为OA期刊资源采集方法需要集成Web查询和自然语言处理两种网页信息资源采集方法的功能。基于网页信息抽取的OA期刊资源采集方法如图1所示。

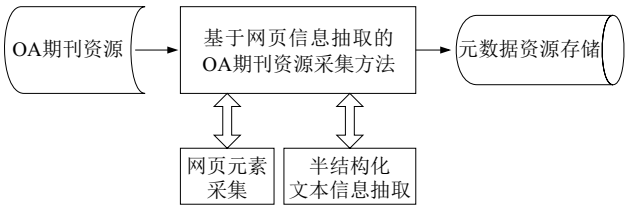

图 1 基于网页信息抽取的OA期刊资源采集方法

网页元素采集指对OA期刊网页中的单一型资源和组合型资源的文本信息进行采集。这些文本信息封装在HTML标签对中，属于网页元素。OA期刊资源采集需要灵活、准确地采集OA期刊网页中的元素。借鉴Web信息采集方法思想，将网页元素采集具体分为三个步骤：首先，将网页解析成DOM树结构；其次，解析出待采集网页元素在DOM树中的路径，并以该路径作为查询条件；最后，使用Web-SQL语句对该网页元素进行查询和采集。该方法对网页依赖度较低，而且不需要大量样本学习，可以灵活应对不同OA期刊网页元素采集。同时，通过待采集元素在DOM树中的路径可以准确定位网页元素位置，保证采集的准确性。

半结构化文本信息抽取指对组合型资源的OA期刊元数据字段进行抽取。组合型资源的文本信息是由多个期刊元数据字段组合而成的半结构化文本。为保证资源采集的全面性，需对组合型资源文本信息中的期刊元数据进行抽取。使用类似基于自然语言处理的信息抽取方法，可以对组合型资源的半结构化文本信息进行抽取。具体步骤为：先对半结构化文本信息进行结构分析，通过人工标注，构建正则表达式对文本进行分解，抽取期刊元数据，进而保证期刊资源的全面采集。

2.4 现有网页信息采集工具特点及适用性分析

为解决OA期刊资源采集的实际问题，同时验证本文提出的基于OA期刊资源网页元数据采集方法的有效性，先对现有3款典型网页信息采集工具进行对比，并对各采集工具在OA期刊网页元数据采集中的适用性进行分析。

2.4.1 现有网页信息采集工具特点分析

国内外3款典型网页信息采集工具对比分析如表3所示。通过对采集工具对比分析发现，3款采集工具都采用类似基于Web查询的网页信息抽取方法，来对网页元素进行采集。不同的是，在实现基于Web查询的网页信息抽取方法时，一部分工具是自动形成定位规则，另一部分工具则需要人工制定定位规则。而对于网页元素中的文本信息，部分采集工具提供正则表达式匹配抽取功能。

2.4.2 网页信息采集工具对OA期刊资源采集的适用性分析

通过上述分析，发现3款采集工具都能准确地采集网页元素，因此，本文主要从采集资源的完整性角度分析各采集工具在OA期刊资源采集上的适用性。本文将OA期刊资源分为单一型资源和组合型资源，本文提出的判断采集工具是否适用于OA期刊资源采集，主要由采集工具是否能对单一型资源和组合型资源进行采集决定。此外，本文在对OA期刊资源采集调研中发现，有超过10%的OA期刊网站存在多套网页模板，即存在网页结构变化的情况。因此，能否对网页结构变化后的资源进行采集也是判断采集工具是否适用于OA期刊资源采集的指标之一。综上所述，单一型资源采集、组合型资源采集以及网页结构变化后资源采集是判断采集工具

表 3 3款国内外典型网页信息采集工具特点对比分析

采集系统	特点分析	优 点	缺 点
八爪鱼采集器	通过内嵌的浏览器与用户交互。用户在内嵌浏览器中点击所需爬取的信息，系统记录选取信息在网页中的唯一标识，并通过标识提取其他网页中与该系列相关的信息	简单的可视化交互，无需用户分析网页结构。可对不同深度的网页进行循环抽取	过渡依赖唯一标识，当网页中唯一标识发生变化，则无法提取信息
火车采集器	主要依赖用户对网页的解析，找到所需提取信息的唯一标识，或找到能分割采集区域的标签字符串，或根据提取信息的网页标签中的规律编写正则表达式来生成抽取规则	可通过正则表达式扩充元数据抽取规则，提高抽取准确度	依赖用户对网页结构的识别，对正则表达式要求较高，无法识别网页结构变化后的页面
Import.io	国外一款较流行的网页信息采集系统。用户可先通过浏览器简单点击，选择其想要抓取的信息；然后通过选择保存数据格式，将采集到的数据保存在本地	交互简单，易操作，对用户要求低，工作量少，适用于列表类网页	对非列表类型采集准确度不高，无法识别网页结构变化后的页面

是否适用于OA期刊资源采集的主要指标。通过对3款工具特点和优缺点分析，结合各工具在OA期刊资源采集上的应用，对3款采集工具的适用性分析如表4所示。

表 4 3款网页信息采集工具适用性对比分析

采集工具	是否能采集单一型资源	是否能采集组合型资源	是否能采集网页结构变化后的资源
八爪鱼采集器	是	是	否
火车采集器	是	是	否
Import.io	是	否	否

通过适用性分析，可以得出两个结论。（1）现有采集工具基本实现了本文提出的采集方法的功能，即对网页元素准确、灵活地采集，对文本信息进行进一步抽取。（2）现有采集工具无法对网页结构变化后的OA期刊资源进行完整采集。由于OA期刊网站存在网页结构发生变化的情况，采集工具不具备网页结构检查功能，形成的采集规则无法对结构变化的网页进行采集。

通过以上分析，虽然现有采集工具基本实现本文提出的基于OA期刊网页信息抽取方法的功能，但并不能对网页结构变化后的OA期刊资源进行有效采集。因此，本文在现有方法基础上，设计一种适用于OA期刊资源采集的系统并进行实证分析，以更好地实现OA期刊资源的全面采集。

3 基于网页信息抽取的OA期刊资源采集系统设计

现有采集工具无法对网页结构发生变化的OA期刊资源进行采集，为全面采集OA期刊资源，进一步验证本文提出的基于网页信息抽取的OA期刊资源采集方法的有效性，在该方法的基础上，还需要提供页面结构检查功能。基于网页信息抽取的OA期刊资源采集框架如图2所示。

基于OA期刊网页元数据抽取的采集框架主要分为数据源、数据采集、数据存储和数据服务四个层次。

数据源层是采集系统面向的数据源。本文主要研究不遵循OAI-PMH协议的OA期刊资源采集方法。根据网页中OA期刊资源的组织形式，为保证OA期刊资源采集的全面和完整，数据源需覆盖结构统一和结构变化两种网页结构的OA期刊资源。

数据采集层是对OA期刊资源实施采集。对于不遵循OAI-PMH协议的资源，主要是在基于OA期刊网页元数据抽取的采集方法基础上，辅以网页结构检查功能，来满足单一型资源、组合型资源以及网页结构发生变化后的期刊资源进行采集。主要解决当前网页信息采集方法无法单独完成OA期刊资源采集，以及当前采集工具无法对网页结构变化后的OA期刊资源采集的问题。

数据存储层主要表现OA期刊资源采集过程中数据的存储过程，包括初始URL、待采集URL和采集规则等的临时存储，以及本地OA期刊元数据数据库等。

数据服务层主要是为采集到的OA期刊资源提供服务，如对采集到的数据进行展示和提供下载服务。

4 实证分析

为进一步验证本文提出的方法，对基于网页信息

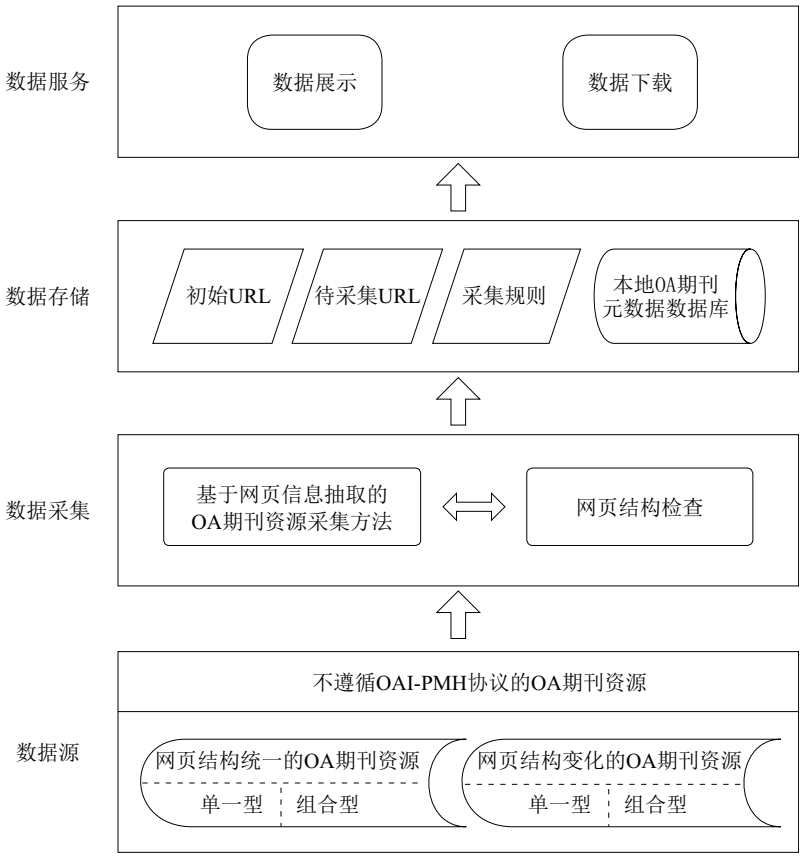

图 2 基于网页信息抽取的OA期刊资源采集框架

抽取的OA期刊资源采集系统的主要功能进行具体的实现。

（1）网页元素采集。使用JavaFX可视化组件Web View，实现资源选择和查询语句自动生成，通过网页解析器Jsoup根据查询语句采集网页信息。具体而言，当Web View组件加载HTML内容时，为每个节点添加事件监听，当鼠标点击某节点时，系统会将该节点赋值给“org.w3c.dom.Node”类型的变量。Node类提供“getParentNode（）”的方法来获取当前节点的父类节点，据此可递归寻找到当前节点到网页根节点的路径。通过将路径中各节点标签名和属性值拼接成Jsoup能够识别的查询语句，再使用Jsoup中select方法对待采集节点的信息进行采集，即可完成网页元素采集工作。

（2）半结构化文本信息抽取。具体实现方式为通过用户标注的分隔符，再根据分隔符位置，提取元数据字段信息。OA期刊网站通常会将“年、卷、期”组合成一条文本信息，如“2017，vol39，no.1”。在抽取具体信息时，先将该条文本信息作为网页元素进行采集，再通过用户在文本中插入分隔符进行标注，将所需采集信息与固定展示信息进行分隔，即“{2017}，vol{39}，no.{1}”“2017”“39”“1”是需要采集的信息，“，vol”“，no.”是固定展示信息。固定展示信息内容通常不会改变，因此，可以根据固定展示信息位置来抽取文本中相应信息。

（3）网页结构检查。根据规定所需采集的必须字段，来作为判断网页结构是否发生变化的标准，如果采集到的必须字段为空则认为当前网页结构已发生变化，需重新选择和采集。如文章标题作为必须字段，在网页元素采集时会判断采集到的该字段是否为空。如果为空则可能有两种情况：一是当前页面确实没有该字段，此页面为脏页面；二是当前页面存在该字段，但该元数据采集规则不适用于当前页面，则可以判断此页面为结构变化后的页面。系统无法识别必须字段为空时属于何种情况，因此，系统会将当前页面加入结构变化页面链接数组中。该轮采集结束后，提取结构变化网页链接数组的第一个链接，在内嵌浏览器中进行展示，由用户对字段为空的情况作出判断。系统对两种情况均提出解决方案，对于第一种脏页情况，直接跳过，

并将该页面链接从结构变化的网页链接数组中删除；对于第二种网页结构变化的情况，用户会在结构变化后的页面上重新进行元数据选择，将形成的新采集规则加入原采集规则集合中，系统会使用新的采集规则继续进行采集。这样往复2—3次便可以遍历网站所有模板，进而采集到全数据，解决OA期刊资源网页结构多变而无法全面采集的问题。

为验证基于网页信息抽取的OA期刊资源采集方法的有效性，本文选择国内外不遵循OAI-PMH协议的10本OA期刊的网站作为采集对象，通过爬虫脚本采集10本OA期刊的论文链接数量，作为采集数量全面性的标准。测试结果如表5所示。

表 5 10本OA期刊资源采集结果对比分析

期刊名称	论文链接数/个	采集到论文总数/篇	网页结构是否发生变化	网页模板数量/个	脏页/个	需用时间/秒
《电子与信息学报》	9 753	9 751	是	2	2	439
《北京航空航天大学学报》	1 497	1 497	是	2	0	342
《电子设计工程》	10 788	10 782	否	1	6	5 858
《工程设计学报》	1 523	1 523	否	1	0	184
《吉林大学学报（工学版）》	4 552	4 472	否	1	80	1 024
《光学精密工程》	7 339	7 339	是	2	0	17 772
《心理学报》	3 902	3 895	否	1	7	94
《冶金分析》	6 030	6 029	否	1	1	3 890
Acta Scientiarum. Animal Sciences	360	360	是	2	0	635
African Journal of Laboratory Medicine	137	137	否	1	0	801

由表5可知，10本期刊共采集到论文45 785篇，采集时间共用31 039秒，其中有4本期刊的网页结构发生变化。通过系统测试结果可以看出，基于网页信息抽取的OA期刊资源采集方法可以灵活应对不同OA期刊资源的采集。在准确率方面，该方法能准确采集单一型资源和文本结构固定的组合型资源，说明其能够适用于OA期刊资源采集工作。基于网页信息抽取的OA期刊资源采集系统的网页结构检查能准确识别网页结构变化，并对结构变化后的资源进行采集。除部分OA期刊网站存在无法访问或无详细信息外，采集到的论文数量与通过爬虫脚本统计到的论文链接数一致。从采集时间上看，平均1 000篇文章的采集时间为678秒。总体而言，基于网页信息抽取的OA期刊资源采集方法，能较好地满足不遵循OAI-PMH协议的OA期刊资源采集需求。

5 总结

本文以OA期刊资源为研究对象，从网页信息采集的角度，对不遵循OAI-PMH协议的OA期刊资源采集进行研究。首先，本文对OA期刊资源特点进行总结，并按照资源在网页中的组织方式将其分为单一型资源和组合型资源；其次，分析对网页采集方法在OA期刊资源采集上的适用性，发现网页采集方法无法单独完成OA期刊资源采集工作。因此，本文提出基于网页信息抽取的OA期刊资源采集方法，该方法综合了网页信息采集方法的功能，不仅能准确、灵活采集OA期刊网页元素，也能对本文信息进行抽取。通过3款典型网页信息采集工具在OA期刊资源采集上的适用性分析，发现各工具均无法对网页结构发生变化的OA期刊资源进行采集。因此，本文对基于网页信息抽取的OA期刊资源采集系统进行设计，增加对网页结构的检查。通过对国内外不遵循OAI-PMH协议的10本期刊网站实证采集，发现4本期刊网站存在网页结构发生变化的情况，并对网页结构变化后的资源进行采集，得到45 785篇论文的元数据信息，证明采集框架能很好地指导不遵循OAI-PMH协议的OA期刊资源采集工作。本文虽然基本满足不遵循OAI-PMH协议的OA期刊资源采集需求，但仍存在如资源采集时间过长等问题，还有待进一步优化。

参考文献

[1] LAENDER A H F,RIBEIRO-NETO B A,SILVA A S D,et al.Abrief survey

of web data extraction tools[J].Acm Sigmod Record,2002,31(2):84-93.

[2] 蒲筱哥.基于Web的信息抽取技术研究综述[J].现代情报,2007,27(10):215-219.

[3] 董娟.基于页面结构分析的网页信息抽取方法研究[D].青岛:中国石油大学(华东),2010.

[4] 于静.基于页面主体提取的WEB信息抽取技术研究[D].南京:南京邮电大学,2013.

[5] 朱江,尚玮姣,姜恩波,等.会议文献开放资源采集与服务系统的建设[J].情报理论与实践,2010(7):117-119.

[6] 王思丽,马建玲,王楠,等.开放知识资源的元数据自动采集策略研究[J].图书馆学研究,2013(12):47-51.

[7] 宋辰.科技情报采集系统的设计及其快速文本聚类方法研究[D].北京:北京工业大学,2014.

[8] 董慧,丁波涛.用OAI-MHP协议解决数字图书馆互操作问题[J].情报科学,2004(6):699-702.

[9] 李勇文.OAI元数据搜索引擎的设计与实现[J].现代图书情报技术,2005(2):37-39,32.

[10] 王芳,王小丽.基于OAI协议的数字档案馆元数据互操作问题研究[J].现代图书情报技术,2007(3):18-24.

[11] 徐方,张静.国内OAI-PMH协议研究综述[J].现代情报,2009(1):89-94.

[12] 郭少友.OAI-PMH元数据的关联数据化方法研究[J].图书情报工作,2011(2):107-111.

[13] 杨东清.开放获取期刊资源库共建共享平台的研究与开发[D].南京:南京农业大学,2010.

作者简介

黄政，男，1992年生，硕士研究生，研究方向：信息资源管理，E-mail：17888802420@163.com。
张学福，男，1966年生，博士，研究员，研究方向：农业知识组织与可视化分析，通讯作者，E-mail：zhangxuefu@caas.cn。

A Research on Open Access Journal Resource Acquisition Method Based on Web Information Extraction

HUANG Zheng, ZHANG XueFu
(Agricultural Information Institute of Chinese Academy of Agricultural Sciences, Beijing 100081, China)

Abstract: Open access journal resources have important academic value, however, some open access journals do not follow the OAI-PMH protocol, and can not collect resources through OAI-PMH protocol. In this paper, based on the characteristics of open Access journal resources, we propose a non OAI-PMH protocol based open access resource acquisition strategy. In this paper, from the point of view of web resources description, this paper summarizes the characteristics of open access journal resources and classifies them from the point of view of web resources description.Based on the applicability of the web information collection method in collecting open access journal resources, this paper proposes a open access journal resource acquisition strategy non based on OAI-PMH protocol, which is based on the method of acquisition open access journal web metadata extraction and design the acquisition system. Through the empirical study of 10 open access journals which do not provide the OAI-PMH protocol at home and abroad, a total of 45 785 papers were collected. It is proved that this method can be effectively applied to the acquisition of such resources. The research enriches the acquisition methods of open access journals, and provides a method to guide the acquisition of open access journals that do not follow the OAI-PMH protocol.

Keywords: Open Access Journal; Open Access Journal Resource Acquisition; Web Information Acquisition; Open Access Journal Resource Acquisition System
（收稿日期：2017-04-14）