开放获取期刊资源采集方法探究

黄政，张学福

（中国农业科学院农业信息研究所，北京 100081）

摘要：为更加有效的采集开放获取期刊资源，本文对开放获取期刊资源进行调研分析，梳理出开放获取期刊资源具有描述粒度细、描述复杂和描述载体结构多变等特点，并根据资源组织形式，将其分为单一型资源和组合型资源。在对比分析各元数据采集方法的基础之上，总结出不同类型的期刊资源适用的采集方法。最后以一个典型的开放获取期刊为对象进行了实证采集分析，结果证明，各类采集方法能有效的采集对应的开放获取期刊资源。

**关键字**：开放获取期刊资源，元数据采集，Web信息采集

# 1.引言

随着开放获取运动的普及，开放获取期刊的数量逐年增加，而期刊资源依然是通过同行评审保证了期刊资源的质量。开放获取期刊遵循免费获取的开放获取原则，但其以不同的组织形式和揭示粒度在网络中进行展示，这为期刊资源采集带来了挑战。研究开放获取期刊资源的采集方法，对开放获取期刊资源个性化的采集以及开放资源平台建设都具有非常重要的意义。

本文首先对专家遴选的开放获取期刊资源进行了调研分析，并对资源的特点进行了梳理总结，同时按照组织形式对资源进行分类。接着对当前元数据采集方法进行了对比分析。最后总结了不同类型开放获取期刊资源对应的采集方法，并进行了实证分析。

# 2.开放获取期刊资源特点与分类

## 2.1 开放获取期刊资源的特点

本文通过对专家遴选的134个国内开放获取期刊网站和15个国外开放获取期刊仓储进行调研，梳理了当前开放获取期刊资源的特点主要有以下几点：

1. 描述粒度细。开放获取期刊资源中的元数据包含了众多字段，如期刊的元数据字段，包括：期刊名、年、卷、期等；论文元数据，包括：题名、其他题名、关键词、英文关键词、摘要、英文摘要、作者、机构、唯一标识号、总页数、起止页、参考文献数量等。相比于其他网络资源，开放获取资源所需采集到元数据更细。
2. 描述复杂。开放获取期刊资源元数据的字段众多，而这些字段通常是以不同的形式展现在网页中，造成了开放获取资源元数据描述的复杂性。主要表现为部分元数据字段在网页中是按照单个字段进行展示，如文章的标题、摘要等是以一个标签单独展示。而部分元数据则是多个字段组合成一条文本信息进行展示，如文章的年卷期等是组合成了一个文本信息进行展示。
3. 描述载体结构多变。在对国内开放获取期刊资源调研过程中发现，部分期刊网站的资源展示页面，在不同时期所采用了不同的网页模板。在结构发生变化的开放获取资源网站中，一般会存在1-3套不等的网页模版。而其他网络资源，如电商平台、论坛等，通常只采用统一的网页模版。

## 2.2 开放获取期刊资源分类

开放获取期刊资源以不同的组织形式在不同的网页中进行展示。通过对国内外开放获取期刊资源的调研和特点总结，根据开放获取期刊资源在网页上的组织形式，将开放获取期刊资源分为：单一型资源和组合型资源。

单一型资源是指网页的一个HTML标签仅展示一个元数据字段信息的资源，该类资源包含的元数据有期刊名称、文章的中英文标题、摘要、关键字、全文获取链接等。此类资源信息揭示简单明了、层次清晰。在对该类资源进行采集时，只需要提取HTML标签对中的信息即可。

组合型资源是指网页的一个HTML标签中封装了多个字段信息的资源，多个字段通常是组合成为一个文本信息进行展示，如期刊的年卷期字段通常是组合成一条文本在网页中进行展示。此类资源信息揭示较复杂。在对该类资源进行采集时，提取HTML标签对中的文本信息之后，还需要进一步对文本进行处理，进而得到各个元数据字段信息。

# 3.元数据采集方法

开放获取期刊资源包含了期刊元数据、论文元数据、论文全文获取链接等，期刊资源的采集重点在于对各元数据的采集，因此本文对元数据采集方法进行分析总结。当前元数据采集方法主要分为两种，一种是基于接口协议的元数据采集方法，即数据提供方和数据采集方遵循统一收割协议，数据提供方按照规定的格式存储数据，并提供数据采集接口，数据采集方调用统一的接口进行元数据的采集。另一种是基于Web信息的元数据采集方法，此方法针对没有统一接口协议，以Web页面为载体来描述元数据。用户通过对Web页面的访问来查看元数据。此类资源采集需结合Web信息采集技术。下面就这两种方法的实现以及在开放获取资源采集中的应用进行讨论，为进一步提出针对开放获取资源元数据采集方法提供指导。

## 3.1基于接口协议的元数据采集方法

基于接口协议的元数据资源采集是对遵循收割协议并提供统一采集接口的元数据进行采集。该方法特点是数据的提供方和采集方都遵守特定的元数据收割协议。其中具有代表性的是OAI-PMH(Open Archives Initiative Protocol for Metadata Harvesting)协议接口[1]。基于OAI协议的元数据获取框架主要是由三部分组成：数据提供方(Data Provider)、服务提供方(Service Provider)以及注册服务器（Registrar Server）。数据提供方负责元数据的生成，即按照统一的规范，在本地建立本地元数据仓库（Repository），并且以OAI的响应(response)向服务提供方发布元数据。服务提供方和数据提供方主要通过6个OAI指令，即Identify、ListMetadataFormat、ListSets、ListIdentifiers、ListRecords、GetRecord， 来对元数据进行收割、整理加工，最终提供增值服务。注册服务器是对服务提供方和数据提供方进行管理的模块。

基于协议的元数据采集特点在于：数据提供方遵循统一的协议，并提供统一的数据采集接口。该方法已十分成熟，能保证元数据采集的准确性与采集效率，但该方法需要数据提供者必须要遵守收割协议，对于网络上存在的大量的非统一接口的数据来源，无法进行采集，具有一定的局限性。

## 3.2基于Web信息的元数据采集方法

基于Web信息的元数据采集方法是指对不提供统一的收割协议的元数据进行采集。元数据由数据提供方自定义描述规则，在Web页面中进行展示。元数据采集方通过对元数据网页进行解析，编写特定抽取规则来抽取其中的元数据。

Web信息采集方法从不同的角度可以将分为不同种类[2]。最常见的分类方式是根据包装器形成方式的不同将网页信息采集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信息抽取、基于HTML页面结构分析的网页信息抽取和基于Web查询的网页信息抽取 [3-8]。表1分别对各种采集方法进行对比，并对其在开放获取资源采集中的应用进行分析。

表 1 各采集方法比较分析

|  |  |  |  |
| --- | --- | --- | --- |
| 方法名称 | 代表系统 | 包装器形成方式 | 适用元数据 |
| 基于自然语言处理的网页信息抽取方法 | 基于WHISK[9]算法的系统 | 依赖自然语言处理技术和语意字典，建立短语之间的关系，所产生的抽取规则通常基于语法约束和语义约束[10]。 | 适用于文本且句子完整的元数据。 |
| 基于本体的网页信息抽取方法 | SNAX[11]系统 | 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。 | 适用于特定主题或领域的元数据。 |
| 基于包装器归纳的信息抽取方法 | WIEN[12]系统 | 根据事先由用户标记的样本实例，应用机器学习方式归纳算法，生成基于定界符的抽取规则。 | 适用于结构清晰，信息较为复杂的元数据。 |
| 基于HTML页面结构分析的网页信息抽取方法 | RoadRunner | 将网页解析成结构树，通过自动或半自动的方式产生抽取规则，定位主题信息，将网页信息抽取转化成对解析树的操作[13]。 | 适用于结构清晰且信息单一的元数据 |
| 基于Web查询的网页信息抽取方法 | Web-OQL[14] | 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的标记，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。 | 适用于信息单一且标识明确的元数据。 |

通过以上对Web信息采集方法的分析，可以从各方法适用元数据的角度将其分大致分为两类：(1)结构化元数据采集方法，包括基于Web查询和基于HTML页面结构分析的网页信息抽取方法。该类方法主要是对结构化网页中封装的信息进行提取，需要用户对网页结构进行分析，找到所需提取元数据在网页中的定位，再通过爬虫工具对元数据字段进行采集。(2)文本类元数据采集方法，包括基于自然语言处理、基于本体和的基于包装器归纳的网页信息抽取方法。该类方法主要对文本信息中包含的元数据进行提取，通过建立语意字典、本体知识库，或者归纳学习生成抽取规则的方法，对一段文本信息中所包含的元数据字段进行提取和采集。

# 4.开放获取期刊资源采集方法与实证分析

针对遵循收割协议的开放获取期刊资源，基于接口协议的元数据采集方法已十分成熟，但具有一定的局限性，因此本文重点讨论不遵循收割协议且没有统一采集接口的开放获取期刊资源。通过对开放获取期刊资源的特点进行分析，根据资源的组织形式将开放获取期刊资源分为单一型资源和组合型资源。而当前基于Web信息元数据采集方法大致可以分为结构化元数据采集方法和文本类元数据采集方法。接下来将结合开放获取期刊资源的特点，分析不同类型的期刊资源适应的采集方法。

对于单一型资源，该类资源的元数据是将单个元数据字段作为一条信息封装在HTML的标签对中。在采集过程中，只需准确采集HTML标签对中的信息即可。因此可以采用结构化元数据采集方法，即基于HTML页面结构分析的网页信息抽取或基于Web查询的网页信息抽取。通过爬虫工具将Web页面解析成HTML语法树，找到元数据所在标签对的标识，按照此标识和标签在文档中的位置编写抽取规则，即可对单一性资源进行采集。

对于组合型资源，该类资源的元数据是多个元数据字段组合成一条文本信息封装在HTML标签对中。在采集过程中，除了将封装在HTML标签对中的文本信息进行采集外，还需要对采集到的文本进行进一步分析，提取文本中包含的元数据字段信息。因此，可以先采用结构化元数据采集方法，即基于HTML页面结构分析或基于Web查询的网页信息抽取方法，将标签对中封装的文本信息进行采集，接着使用文本类元数据采集方法，即基于自然语言处理或基于本体或基于包装器归纳方式的信息抽取方法，对文本信息进行分解和提取，进而得到最终需要的元数据字段。

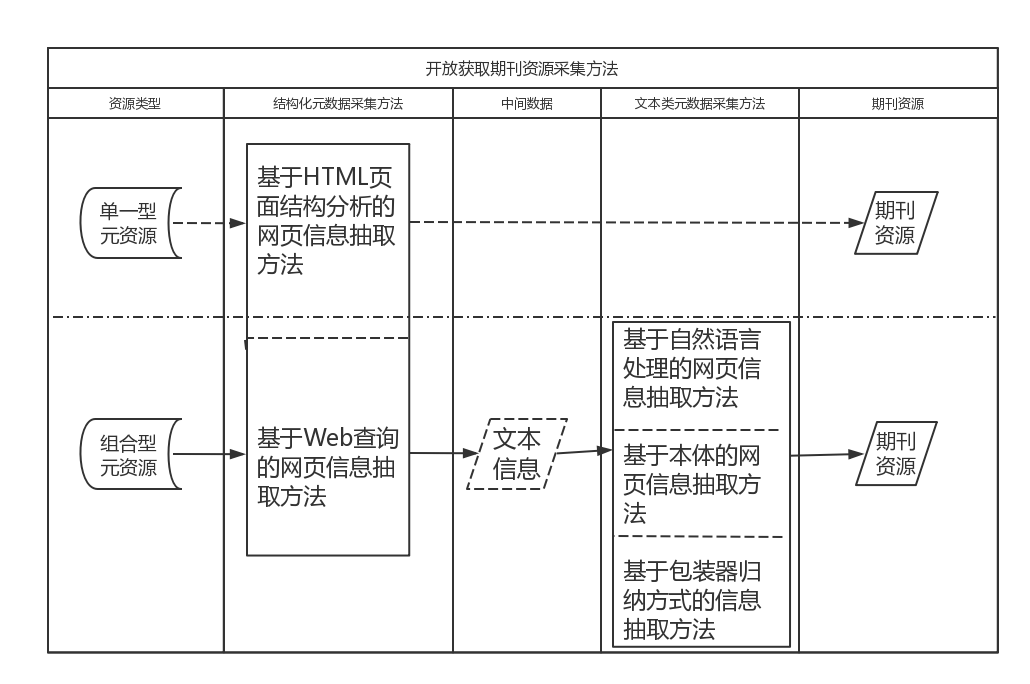


图 1开放获取期刊资源采集方法

为验证各类型开放获取期刊资源对应采集方法的有效性，本文在专家遴选的开放获取期刊中选择了其中一个具有代表性的开放获取期刊进行验证。该开放获取期刊网页同时包含了单一型资源和组合型资源。单一型资源包括：期刊名称、论文标题、论文英文标题、论文摘要、英文摘要、关键词、英文关键词、全文下载链接，DOI；组合型资源包括：期刊年卷期、论文作者、论文作者机构、论文开始页、论文结束页等。

对于单一型资源，本文使用基于Web查询的网页信息抽取方法进行元数据字段提取，具体是使用java语言开发的第三方网页解析工具Jsoup将网页解析成HTML树，然后分析网页源代码，找到封装元数据的HTML标签对的标识，编写类似于SQL的查询语句作为抽取规则，进而对标签对中的元数据进行提取。比如对于文章的标题，在HTML树中的唯一标识是”span class="J\_biaoti"”，编写查询语句select(apan[class=J\_biaoti])，对标题元数据进行采集。

对于组合型资源，先通过基于Web查询的网页信息抽取方法将组合的文本信息进行采集，接着使用基于包装器归纳方式的信息抽取方法对文本信息进行进一步解析，进而提取出各元数据字段。如文章的年卷期通常是以类似于“2016, Vol.38 No.2”这样的文本进行展示。使用基于包装器归纳方式的信息抽取方法，制定基于定界符的抽取规则，如“年”字段信息是文本初始位置到符号“，”之间的信息，“期”字段信息是字符串“Vol.”和字符串“No”之间的信息，“卷”字段信息是字符串“No.”到文本结尾之间的信息。基于定界符的抽取规则，即可从文本信息中提取“年”“卷”“期”三个元数据字段的信息。

基于以上的策略，本文对该开放获取期刊的资源进行了采集，最终采集到了该期刊1万多篇文献资源。采集的数据的准确性和全面性都较高，证明了针对不同类型的元数据的采集方法具有一定的可行性。

# 5.总结

本文首先对专家遴选的开放获取期刊资源的特点进行调研，根据资源的组织形式将开放获取资源资源分为单一型资源和组合型资源。并在对现有的元数据采集方法进行了对比分析的基础之上，提出了不同类型的开放获取期刊资源应该采用的采集方法。最后通过选择一个典型的开放获取期刊做了实证研究，结果证明，各类期刊资源对应的采集方法能有效的对资源采集，采集的准确率和全面性比较高。本研究虽然基本上满足了开放获取期刊资源的采集需求，但仍存在一些问题，如采集所需时间较长等，还有待进一步优化。

参考文献

[1] 徐方,张静. 国内OAI-PMH协议研究综述[J]. 现代情报,2009,01:89-94.

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

[3] Alberto H. F. Laender,Berthier A. Ribeiro-Neto,Altigran Soares da Silva,Juliana S. Teixeira. A Brief Survey of Web Data Extraction Tools.[J]. SIGMOD Record,2002,31:.

[4] 轩艳艳. 基于XML的Web信息抽取研究与实现[D].武汉理工大学,2008.

[5] 贺智平. Web信息自动抽取技术研究[D].西安电子科技大学,2006.

[6] 李猛. 基于DOM的Web信息抽取技术的研究与实现[D].大连理工大学,2008.

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

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

[9] Stephen Soderland. Learning Information Extraction Rules for Semi-Structured and Free Text[J]. Machine Learning,1999,341.

[10] 吴晓彦. 基于结构语义熵的互联网商品信息抽取技术研究[D].复旦大学,2009.

[11] 拜战胜. SNAX系统数据采集研究[D].中南大学,2009.

[12] Kushmerick N. Wrapper Induction for Information Extraction [D]. University of Washington, 1997.

[13] 杜鹏. 基于视觉特征的WEB页面信息抽取技术的研究[D].西北师范大学,2009.

[14] Arocena G.O, Mendelzon A O.WebOQL: Restructuring Documents Databases and Webs. Proceedings of the 14th IEEE International Conference on Data Engineering ,Orlando,Florida,1998:24-33.

作者简介

黄政，男，1992年生，中国农业科学院农业信息研究所硕士研究生，研究方向:信息资源管理，E-mail:17888802420@163.com。

张学福，男，1966年生，中国农业科学院农业信息研究所研究员，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。

通信地址:中国农业科学院农业信息研究所5309科室，北京，100081

联系方式:17888802420

E-mail：[17888802420@163.com](mailto:17888802420@163.com)

Open Access Journal Resources Extraction Method Research

**Abstract:** In order to complete the extraction of open access journal resources more effectively, we do the research on open access journal resources selected by experts, and sort out the open access resources metadata have some characteristics, such as descript detail, descript complex and description vector structure variable. According to the organization form of the resources, we divide the metadata into the single resources and the combination resources. On the basis of the comparative analysis of the metadata extraction methods, the extraction methods of different types of resources are discussed. Finally, we use a typical open access journal for empirical analysis, results show that the extraction method of different kinds of resources can work effectively.

**Keywords**: Open Access Journal Resource, Metadata Extraction, Web Information Extraction