基于页面结构检查的开放获取资源元数据采集方法研究

摘要：为更加有效的完成开放获取资源采集，本文通过对专家遴选的开放获取资源进行调研分析，总结了开放获取资源的特点。并通过对当前元数据采集方法和采集系统的特点以及在开放获取资源元数据采集上的应用进行了对比分析，总结出当前的方法和系统主要存在两个问题：(1)局限性；(2)数据采集不全面。最终结合开放获取资源元数据的特点，研究提出了基于页面结构检查的开放获取资源元数据采集方法，并实践证明该方法能有效满足开放获取资源采集需求。

**关键字**：开放获取资源，元数据采集，Web信息采集

Open Access Resource Metadata Extraction Method Research

**Abstract:** In order to complete the extraction of open access resources metadata more effectively, we do the research on open access resources selected by experts, and sort out the open access resources metadata have some characteristics, such as metadata descript more detail, metadata description complex and focus more on data quality. Through the comparison of current methods and systems of metadata extraction, and analysis the application of method and system on open access resource, we summarize that there are two main problems: (1) limitations; (2) data collection is not comprehensive. Then we put forward a new extraction method which based on web structure checking mechanism of open access resource metadata. And practice shows that the method can effectively meet the requirements of open access resource extraction.

**Keywords**: Open Access Resource; Metadata Extraction; Web Information Extraction;

# 1.引言

随着开放获取运动的普及，开放获取资源的价值逐步逼近主流商业学术资源。开放获取资源通常是以不同的组织形式和元数据揭示粒度分散在各种维度层次的网络中，而且元数据采集的需求也各不相同，这为元数据采集提出来新的挑战，同时为数字图书馆建设带来了新的需求。研究开放获取资源的采集方法，对满足个性化的开放获取资源采集需求以及开放资源平台建设都具有非常重要的意义。

本文首先对开放获取资源的分类和特点进行了调研分析，接着结合开放获取资源的特点对当前的资源采集方法和采集工具进行了对比，总结出当前采集方法和采集工具在应用于开放获取资源采集时的不足，最终提出一套针对于开放获取资源采集的方法。实验证明，此方法切实可行，对开放获取资源元数据采集具有一定的帮助作用。

# 2.开放获取资源分类和特点

## 2.1 开放获取资源分类

随着互联网的发展，开放获取资源形成了不同类型。目前最为常见的是开放获取仓储(Open Access Archives)和开放获取期刊(Open Access Journals)两种类型[1]。

开放获取仓储(Open Access Archives)是遵循开放获取原则的一种电子版科技文献存储和检索的数据库。研究学者们通过向开放获取仓储提交已发表或未发表的研究成果，使得成果能迅速被其他学者学习并讨论，从而促进当前学科的发展。开放获取仓储中可以包括：电子文档、论文、课程资料、数据文件、声像文件、机构记录以及任何类型的数字文档。开放获取仓储主要分为两种类型：一种是收集某一特定学科或主题的学科知识库，另一种是由多个学术机构共同组成的机构知识库[2]。

开放获取期刊(Open Access Journals)是遵循开放获取原则的期刊。开放获取期刊经过同行评审，并通过网络载体，对期刊资源进行描述。开放获取期刊与传统期刊的区别不在于期刊的载体是纸本还是电子，而在于对期刊的访问方式和访问权限。所有的用户可以不受价格和权限的限制免费从互联网上获取此类期刊的学术信息[3]。

## 2.2 开放获取资源元数据的特点

本文通过对专家遴选的134个国内开放获取期刊网站和15个国外开放获取仓储进行调研，发现开放获取资源元数据比其他网络资源元数据有一定的区别。主要表现为：

1. 开放获取资源元数据描述粒度细。《NSTL开放资源描述元数据标准规范》中对开放获取元数据的描述定义了7类元素集：开放资源集合、开放资源作品、责任者、机构、开放资源使用权益、获取来源、管理信息，而这7类元素集又包含了众多元数据字段，如题名、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、年、卷、期、期刊名称、全文下载链接等。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. 开放获取资源元数据描述复杂。开放获取资源元数据描述复杂主要表现在两个方面：一是网页中元数据展示多样，即不同开放获取资源的平台网站和期刊网页所揭示的元数据薄厚不均，二是描述元数据的网页结构多变，即在一个期刊网站中，网站不同时期所采用的网页模板不同。
3. 开放获取资源注重元数据的质量。开放获取的元数据除了采集和发布之外，还要作为其他研究的基础，如知识演化分析、学科热点分析等都需要论文的关键词等元数据作为数据基础。开放获取资源的元数据的质量直接影响到其他研究的效果。因此开放获取资源对元数据的质量要求比较高。

# 3.元数据采集方法与系统

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的特点在于，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[4]。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。

数据提供方与服务提供方的交互过程为：服务提供方向数据提供者发起一个HTTP请求；数据提供方对此HTTP请求进行解析，判断是否为OAI-PMH协议中的六个有效请求之一；若为有效请求，数据提供方根据请求进行相应，返回元数据结果，若不是有效请求则返回错误信息。

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，提高了元数据采集收割的效率。但不足之处是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信息的元数据资源采集方法

基于Web信息的元数据指的是不提供统一的元数据收割协议，由数据提供方自定义描述规则，在Web页面中进行展示的元数据。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。本文对当前主要的Web信息采集方法和Web信息采集系统进行了调研分析，并对各方法和系统在开放获取资源元数据采集上的应用进行了对比分析。

### 3.2.1 Web信息采集方法分类

Web信息采集方法从不同的角度可以将分为不同种类[5]。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、基于Web查询的网页信息抽取和基于页面视觉特征的网页信息抽取[6-11]。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 | 包装器形成方式 | 适用领域 | 应用于开放获取资源采集 |
| 基于自然语言处理的网页信息抽取 | 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[12]。 | 含有大量文本且句子完整的网页 | 开放获取资源元数据描述是粒度化的，并不是用一条文本来展示。 |
| 基于本体的网页信息抽取 | 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，保留相关信息。 | 特定主题或领域的网页 | 需领域专家构造完整的本体知识库，需要专家投入相当多的时间。 |
| 基于包装器归纳的信息抽取 | 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。 | 具有相同结构和编码的网页。 | 开放获取资源页面结构多变，需提前搜集所有样本。 |
| 基于HTML页面结构分析的网页信息抽取 | 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]。 | 由共同模板生成，具有统一结构的网页。 | 要使用了大量复杂的启发式搜索算法来形成抽取规则。 |
| 基于Web查询的网页信息抽取 | 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。 | 网页结构中抽取字段有清晰的唯一标识的网页。 | 需要人工对网页进行识别网页结构，对于结构变化的网页无法采集。 |
| 基于页面视觉特征的网页信息抽取 | 利用Web信息的视觉特征，例如字体的大小等信息，将一个网页分割成多个不同的视觉信息块，删除与主题无关的信息之后，再使用包装器对信息进行抽取。 | 主要适用于网页结构清晰的网页。 | 由于视觉特征具有复杂而又不确定的特性，使用起来很难保证规则集的一致性[14]。 |

各种采集方法在对结构多样的开放获取资源采集具有一定的不适应性。基于自然语言处理和基于本体的网页信息抽取事先需要领域专家的参与，编写语义字典或本体知识库，更加适用于对网页中相关主题信息的提取。基于Web查询和基于HTML页面结构分析的网页信息抽取需要用户对网页结构进行分析，对用户要求较高，且得到的抽取规则无法适用结构变化的页面，适用于结构统一的网页。基于页面视觉特征的网页信息抽取适用于结构清晰的网页，而且往往还需结合其他方法对元数据进行采集，适用范围有限。基于包装器归纳方式的信息抽取通过样本页面归纳学习抽取规则，但样本网页有时无法覆盖到所有网页的描述规则，训练所得到抽取规则无法抽取网页结构变化后的元数据信息。

### 3.2.2 Web信息采集系统比较

随着网络环境的开放和网络资源的增长，对网络资源采集的需求催生了一些商用的Web信息采集系统。这些系统通常用于对新闻网站、电商平台、论坛等结构统一的网站信息进行采集。表2是本文对当前国内外主流采集系统进行的比较，并对各系统在开放获取资源采集上的应用进行分析。

表 2 国内外典型网页信息采集系统比较和分析

|  |  |  |  |
| --- | --- | --- | --- |
| 采集系统 | 简介 | 优点 | 应用于开放获取资源采集 |
| 八爪鱼采集器[15] | 通过内嵌浏览器与用户进行交互，用户在内嵌浏览器中点击所需抓取的信息，系统记录选取信息的路径，进而形成抽取规则来提取采集页中的信息。 | 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。 | 网页中唯一标识发生变化，则无法提取信息。 |
| 火车采集器[16] | 依赖于用户对网页的解析，通过用户查看网页源代码，找到所需提取信息的唯一标识，或者找到能分割采集区域的标签字符串，或者根据所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的抽取规则，来应对网页中唯一标识变化的情况，提高抽取的准确度。支持丰富的拓展。 | 依赖用户对网页结构的识别，对正则表达式要求较高。 |
| ET采集器[17] | 需要用户对网页源代码进行分析，标记出采集的字段，系统生成类似正则表达式的抽取规则。通过内置的测试工具能快速的查看当前抽取规则的抽取结果，并反复调整抽取规则，来获取最佳的抽取规则。 | 通过自动生成类似于正则表达式的抽取规则，可提高抽取的准确度，内置的测试工具可以快速的查看抽取结果。 | 需要用户反复对网页源代码进行分析。 |
| 网络矿工采集器[18] | 将用户所需采集的网页分为导航页和采集页，需要用户解析网页源码，通过设置所需提取信息的起始位置和终止位置，来设定类似正则表达式的抽取规则。 | 类似正则表达式的抽取规则使采集准确度高，系统提供大量的按钮和提示，减少了用户使用的难度。 | 需要用户通过识别网页结构寻找抽取信息的起始位置。 |
| Import.io[19] | 国外一款较流行的网页信息采集系统。用户在浏览器中输入链接后，通过简单的点击操作，选择所需抓取的信息，然后选择保存数据的格式，完成数据采集。 | 交互简单，易操作，对用户要求非常低，工作量少，适用于结构统一的列表类的网页。 | 多数开放获取资源网页是非列表类型，采集不够全面。 |

从系统易用性的角度来看，Import.io和八爪鱼采集器提供内嵌的浏览器，用户只需在浏览器中点击所需采集的信息就能自动生成抽取规则，对用户要求较低。而其他系统有的需要用户分析网页，有的需要编写正则表达式等，需要用户有一定的专业知识背景才能生成有效的抽取规则，对于普通用户的使用有些难度。

从数据全面性的角度来看，本文通过对某一期刊网站进行试验发现，对于单个页面或者多个具有相同编码特征时，如标题的HTML标签中的唯一标识都为“J\_biaoti”，采集器通过编写的采集规则能全面采集元数据。但当此期刊网页中的编码特征变化，如标题的HTML标签中的唯一标识由原来的“J\_biaoti”变成了“J\_ZhaiYao\_BiaoTi”，各采集器都无法对结构变化后的信息进行采集。因此数据采集的全面性与网页结构的统一性有非常大的关系。当网页结构统一，各采集器的准确率会非常高，但是网页结构不统一，则需对不统一的网页重新编写采集规则。

# 4.开放获取资源采集框架

通过上述讨论，不管是基于接口协议的元数据采集方法，还是基于Web信息的元数据采集方法，在应用于开放获取资源采集上都存在一些需要改进的地方，主要表现为：

(1) 采集方法具有局限性。基于接口协议的元数据采集方法，受限于资源的存储与收割必须遵循相关协议，使得此方法的使用范围局限于遵循协议的机构，而对于协议之外的元数据，就无法使用此方法进行采集。

(2) 采集数据不全面。主要原因在于开放获取资源元数据描述方式多样。当前基于Web信息的元数据采集方法和采集系统仅仅通过一个或者少量几个样本网页就形成了采集规则，而开放获取资源元数据描述网页的结构是经常变化的，这种情况下形成的采集规则往往是不全面的，进而无法采集到所有的元数据记录，造成采集的准确度不高。

为寻找出一套能适用于开放获取资源元数据采集的策略，本文分别对以上两个问题进行解决方法的讨论。

(1) 在局限性方面，开放获取资源最终都是以网页形式对元数据进行展示。因此为突破协议的约束，可以从网页信息采集的角度入手，研究一套能够针对开放获取资源特点的网页信息采集方案，来应对所有在网络上展现的开放获取资源，而不受协议采集的约束。

(2) 在数据全面性方面，由于开放获取资源的网页结构多变，为得到全面的采集规则，人工识别网页结构的变化不切实际。但可以将网页结构变化检测交给计算机，即如果已形成的采集规则不适合当前页面，说明页面结构发生了变化，然后返回给用户，让用户重新选择，如此循环来获得全面的采集规则，进而保证采集数据的全面性。

此外，不同的采集者对开放获取资源元数据采集需求可能不同，因此，采集到字段应该由用户自己定义。用户可以通过交互式的界面，根据采集需求对网站中的元数据进行选择，并由计算机记录下当前字段的采集路径，进而形成当前资源的采集规则。

根据以上几点的讨论，本文结合当前采集方法、采集系统以及开放获取资源的特点，提出基于页面结构检查的开放获取资源元数据采集方法，具体流程如下图所示：

|  |  |
| --- | --- |
| 用户交互模块 | 点击需要采集元数据  输入开放获取资源链接  系统形成抽取规则 |
| 检测模块 | 循环检测网站其他采集结构是否变化  增加当前抽取规则到规则库  否  是  是  否  检测抽取规则是否已经存在 |
| 元数据采集与存储模块 | 元数据存储  信息抽取 |

图 1基于页面结构检查机制的开放获取资源采集方法

采集方法主要分为三个模块：用户交互模块、检测模块和元数据采集与存储模块。

用户交互模块，通过内嵌浏览器与用户进行交互。在此模块中用户在内嵌浏览器的网页中点击所需提取字段，无需人工解析网页源代码，后台自动形成抽取规则。不仅降低了对用户的要求和工作量，而且通过用户自主的选择所需提取的字段，使得采集信息的准确性和可拓展性都得以保证，而且使用于任何形式的网页结构，通用性得到增强。

检测模块分为采集规则检测和网页结构检测两块。采集规则检测主要是对当前采集的URL和抽取规则在抽取规则库中进行检查，如果存在，则证明该资源已经采集过，就直接从数据库中将之前采集到的元数据根据用户需求返回给用户。如果不存在，则说明该资源未经过采集，将当前采集，就进入到网页结构检测。网页结构检测是对网站中所有需要采集的网页进行循环检测，查看是否有网页结构发生变化。若没有就进行网页元数据的采集。若有变化则再进入用户交互模块，让用户重新选择，形成新的采集规则，再进入检测模块。此模块一方面解决了开放获取资源网页结构变化的问题，提高了采集的覆盖度，另一方面对已采集到数据不用再进行重复采集，使得资源能重复利用。

元数据采集与存储模块是指根据当前的采集规则库中的采集规则，使用包装器归纳法对当前资源进行采集。采用类似于当前较流行的采集系统中所使用的分布式采集和存储策略，提高采集和存储的速度。

此方法结合了当前流行的采集系统与用户交互的特点，不需要用户解析网页，降低了对用户的要求。而重要的是增加了检测模块，通过自动检测采集网站的网页页面结构和提取记录，解决了开放获取资源结构变化多样的问题，提高了开放获取资源采集的效率。

# 5.总结

开放获取资源的数量和质量逐步提升，元数据的利用价值更加丰富，使得开放获取资源元数据采集的研究具有重要意义。本文首先对开放获取资源的特点进行调研，接着对现有采集方法和系统在开放获取资源采集上的应用进行了对比分析，总结出开放获取资源采集当期所面临的主要问题是采集方法局限、采集数据不全面等问题。并在此基础之上，提出基于页面结构检查的开放获取资源元数据采集方法。该方法适用于所有在网页上进行展示的开放获取资源的元数据采集，打破了基于接口协议的元数据采集方法只适用于遵守协议的机构之间采集的局限性。框架通过用户交互模块满足用户个性化的采集需求，通过检测模块，能检测出结构变化的网页，进而根据变化的网页来完善抽取规则，解决开放获取资源网页结构多变的问题。最后通过实践，使用该策略采集了134个国内开放获取期刊近450000篇文献元数据。采集的准确率和全面性比较高。该框架虽然基本上满足了开放获取资源的采集需求，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]王丹. 开放获取仓储发展研究[D].东北师范大学,2011.

[2]韩月萍. 网络开放存取的学术资源及其获取策略[J]. 青海师范大学学报(哲学社会科学版),2010,05:163-166.

[3]李武. 开放存取出版的两种主要实现途径[J]. 大学图书馆学报,2005,04:58-63.

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

[5]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.

[6]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:.

[7]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.

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

[9]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.

[10]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.

[11]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.

[12]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.

[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]张伟. 基于视觉特征的Web信息抽取技术的研究与实现[D].华东师范大学,2008.

[15]八爪鱼采集器[EB/OL].[2016-04-10]. <http://www.bazhuayu.com/about#about-us>

[16]火车采集器[EB/OL].[2016-04-10]. <http://www.locoy.com/product>

[17]ET采集器[EB/OL].[2016-04-10]. <http://www.zzcity.net/>

[18]网络矿工采集器[EB/OL].[2016-04-10]. http://www.minerspider.com/product/index.aspx

[19]Import.io[EB/OL].[2016-04-10]. https://www.import.io/

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：591701477@qq.com