元数据自动抽取研究新进展*

曾 苏[1,2] 马建霞[1] 张秀秀[1,]

[1]（中国科学院国家科学图书馆兰州分馆 兰州 730000）

[2]（中国科学院研究生院 北京 100190）

[摘要] 分析了元数据自动抽取的现实需求，对元数据自动抽取的相关研究进行了阐述，然后对 DROID、 NLNZ Metadata Extractor、Metadata Miner Catalogue PRO 三种典型的元数据自动抽取器进行了分析比较；在提出目前元数据自动抽取技术局限性的基础上，对该技术作了总结和展望。

[关键词] 元数据；自动抽取； 抽取器

[分类号] G250.76

New Development of Automatic Metadata Extraction

Zeng Su

(Lanzhou Branch, National Science Library, Chinese Academy of Sciences, Lanzhou 730000)

(Graduate University of Chinese Academy of Sciences, Beijing 100049)

Ma Jianxia

(Lanzhou Branch, National Science Library, Chinese Academy of Sciences, Lanzhou 730000)

[Abstract] This paper analyses realistic demands of automatic metadata extraction, elaborates related research on automatic metadata extraction and compares three typical automatic extractors of metadata: DROID, NLNZ Metadata Extractor and Metadata Miner Catalogue PRO. On the basis of discussing present limitations of automatic metadata extraction, the article gives a summary and prediction of this technology.

[Keywords] metadata; automatic extraction; extractor

1 元数据自动抽取的现实需求

随着当代信息技术的飞速发展，以印刷型书刊资料为主要收藏载体的传统图书馆逐渐难以适应信息社会不断增长的信息需求，E-only、E-first 为主要特征的数字图书馆必将成为科研人员的主要信息源。元数据为数字图书馆的信息单元和数据集合提供规范、普遍的描述方法和检索工具，并且为其分布、异类资源的信息体系提供整合的工具与纽带。元数据对数字图书馆而言至关重要，离开了元数据的数字图书馆将无法提供有效服务。

然而面对海量文献描述的需要，如何快速、高效产生元数据成为数字图书馆建设过程中的一大难题。元数据主要有人工输入和自动生成两种方式，人工输入又可分为作者、信息加工人员两种。当前数字图书馆建设过程中，由于没有规定作者必须提交文档的元数据信息，元数据大部分由图书馆员逐条输入。这不仅花费了大量的人力、物力和时间，而且也越来越不能满足海量文献描述的需要。若元数据可以自动生成、自动抽取，必将大大减轻信息人员的的工作负担和极大地提高工作效率，而且可以避免元数据人工录入过程中的主观性和不准确性。

2 元数据自动抽取相关研究

目前国内外学者对元数据自动抽取已有不少研究，主要可分为以下几类：

*本文系国家社会科学基金项目“机构知识库建设与应用研究”(项目编号：07BTQ019)的研究成果之一。

2.1 对特定格式文档的元数据自动抽取

数字图书馆中文档类型主要为PDF、DOC、PPT、HTML、JPEG等格式。由于PDF是数字图书馆中最常见的存储格式，并且其对各种设备输出结果的兼容性，对PDF格式文档进行元数据自动抽取的研究最多，现有的元数据抽取器都能实现对PDF文档的自动抽取。其他格式文档的元数据自动抽取也有相关研究，如DC.dot[1]可为Word和PowerPoint文档自动生成元数据。

2.2 对不同类型元数据的自动抽取

文档的元数据主要包括题名信息、作者信息、来源信息、关键词信息、摘要信息、引文信息、外部特征信息等。文献[2]介绍了一种表格搜索引擎——Tableseer，表格元数据抽取器是此搜索引擎的一部分，采用文本信息剥离器和具有分箱功能的表格探测器实现对表格环境/地理元数据、表格框架元数据、表格附属信息元数据、表格布局元数据、表格单元格内容元数据及表格单元格类型元数据的自动抽取。Min-Yuh Day等人[3]在其文章中介绍了基于等级知识描述框架的INFOMAP方法实现对引文元数据的自动抽取，如作者、标题、期刊、卷期号、出版年和页码信息；Eli Cortez等人[4]提出了基于知识的引文元数据抽取方法，可从任意格式的文档中抽取准确的元数据。西安交通大学胡云华等人在其文章[5]中指出利用机器学习模型（主要采用字体特征等作为格式化信息），从Word、Powerpoint文档中自动抽取题名元数据信息。

2.3 对 Web 站点元数据的自动抽取

贺亚锋[6]介绍了两种Web站点元数据自动生成工具，英国R0ADS计划元数据编辑器和澳大利亚MeatWeb计划的元数据生成器。薛叶伟、胡云华[7]等人采用机器学习的方法，从HTML网页自动抽取文章标题元数据。

2.4 对中文文献元数据的自动抽取

北京理工大学于江德[8]介绍了基于CRFs（Conditional Random Fields，条件随机场）算法的论文元数据抽取方法，并对中文和英文论文的元数据抽取结果进行实证研究，得出该方法可有效地实现从中英文论文中抽取作者、题名、期刊、卷期号、出版年、页码等元数据信息。李朝光、张铭[9]等人在不采用语法分析等复杂的自然语言处理手段的情况下，利用正则表达式抽取论文的页眉信息、文章标题、作者信息、摘要信息、关键词信息、引文信息。这种元数据抽取方法只能针对论文文献进行抽取，而且仅限定于PDF格式论文的抽取，对其他格式、其他类型的的文档进行元数据自动抽取则不能完成。

3 典型元数据自动抽取器比较

目前存在的元数据自动抽取器主要有：英国国家档案馆的 DROID 文件格式辨别工具、新西兰国家图书馆的 Metadata Extractor 软件和法国的 Metadata Miner Catalogue PRO 软件。DROID 和 Metadata Extractor 为开源软件，用户可以在网上免费下载使用并可对其进行二次开发；而 Metadata Miner Catalogue PRO 则需付费使用，试用版仅能处理十条数据。

3.1 DROID

DROID[10]（Digital Record Object Identification）是 2005 年由英国国家档案馆

数字资源长期保存小组开发的，能实现对批量文件格式的自动识别，其目的是满足任何数字知识库准确识别所存储数字对象格式的基本需要。DROID的各种版本（最新版本为V2.0）可在网上免费下载，可分别在Windowes、Mac OS X系统下安装运行。DROID的运行环境：Windows 2000、 XP、Vista 或Macintosh OS X操作系统，最小 512M内存；预设JAVA运行环境，Sun JRE1.5 或更新的版本；与因特网连接，以便于签名文件的自动更新。

DROID 是基于 JAVA 语言的可跨平台操作的工具，提供 API 接口，可简单与其他系统进行整合。用户使用 DROID 软件可以选择图形界面和命令行界面两种方式。此工具操作简单，可选择添加单个文件和整个文件夹的方式进行处理，只要几步即可完成对数字文档的识别；识别速度快，可批量实现对电子文档的识别，在很短的时间内即可完成；处理结果可选择 XML、CSV 格式存档，还可以进行打印和输出结果预览；中英文文档都能被识别，可满足处理中文文献的需要。

目前，DROID 仅能实现对文件 PUID（PRONOM 唯一标识符）、MIME Type（资源的媒体类型）、Format（格式）、Version（签名版本）、Status（状态说明）、Warning（警告信息）的识别。从识别的结果看，DROID 只能对电子文档的外部特征进行识别，对其内容特征如作者、时间等元数据则无法自动抽取。并且，DROID 不能对所有类型电子文档的外部特征进行识别，对 RM 等格式则不能识别。DROID 的功能是不断完善和发展的，今后会添加对软件类型、硬件环境、压缩算法和字符编码机制的扩展识别。

3.2 NLNZ Metadata Extractor

NLNZ Metadata Extractor[11] （简称Metadata Extractor）是由Sytec Resources为新西兰国家图书馆开发的，主要用于处理数字化文档和提取元数据信息。这个软件开发于2003 年，现在已经更新到V3.0 版，从 2007 年开始可免费下载使用。Metadata Extractor可以在Windows和Linux系统环境下运行，必须预设JAVA运行环境。它可以对各种类型电子文档进行元数据抽取，包括图像文件（BMP、GIF、JPEG和TIFF格式）、办公文档（MS Word2.6、Word Perfect、Open Office 、MS Works、MS Excel、MS PowerPoint和PDF格式）、音频视频文件（WAV和MP3 格式）、标记语言文档（HTML和XML格式），提供Native form、NLNZ Data Dictionary两种输出格式。

Metadata Extractor 采取“Extract in Native form”、“NLNZ Data Dictionary”两种不同的抽取格式，会产生不同格式的输出结果：Native form 与 nlnz_presmet.xsd 。Native form 是按 XML-DTD 格式描述的，抽取的元数据信息是可获得的关于电子文档的信息，主要包括 object 相关信息（名称、ID）、抽取主机系统时间（日期、时间）、结构类型（硬件环境、软件环境、抽取完成者）、电子文档相关信息（文档的保存路径、名称、类型、大小、时间属性、软件版本信息）； nlnz_presmet.xsd 是按 XML Schema 格式描述的，这是新西兰国家图书馆主要采用的格式，主要包括以下字段：元数据项（文件名、URL、URI、文件类型、修改时间等）、文件类型元数据（软件相关 ID 信息、开发商、版本、加密算法等相关信息）。

Metadata Extractor对硬件环境要求低，在一般的PC上都可运行（需一定的软件配置），运行时占用内存少，响应速度快；对元数据信息可进行批量抽取，在一定程度上减少了用户和图书馆员手工录入元数据的负担；抽取的结果以XML形式保存，能直接导入到元数据保存仓储[12]和机构知识库中。但该抽取器从电子文档的头文件中抽取元数据信息，不能对电子文档全文进行抽取。抽取的字段大都是电子文档的外部特征，如文档名称、类型、修改时间、URL、软件版本等，重要的内容特征如题名、作者、文摘、引文等字段还不能抽取；并且对中文文献不能进行有效的元数据抽取，文件名中中文字符无法显示，只能显示英文和数字部分。

3.3 Metadata Miner Catalogue PRO

Metadata Miner Catalogue PRO[13]（简称Catalogue）是由Soft Experience开发的的商业软件，主要可用于抽取题名、作者、主题、关键词等描述性元数据信息。Catalogue可从Microsoft Office、OpenOffice、StarOffice 、Visio documents 、HTML、PDF、JPEG、Tiff、PSD文档中实现元数据的自动抽取，提供英语、法语、德语、葡萄牙语、西班牙语 5 种界面语言。 Catalogue主要有以下功能[14]：

（1）收集、提取元数据并形成目录文件，便于管理元数据信息。主要包括以下元数据信息：Microsoft Office（PPT、Word、Excel）、OpenOffice、StarOffice 的特征元数据识别，包括文档类别、题名、作者、主题、关键词、页数、段落数、行数及创作修改时间等元数据；HTML 网页的关键词分析，包括 HTML 文件<title> </title>标签和<meta>标签，按 DC schema 进行元数据抽取；PDF 文档的版本、作者、创作及修改时间、题名、主题、关键词、页数等元数据信息的抽取；IPTC（国际报业电信委员会）的 JPEG/TIFF/PSD 格式图象元数据的抽取，可抽取出名称、编辑状态、关键词、日期、标题等元数据信息；Adobe XMP（可扩展元数据平台）的元数据的抽取，可实现对最新 Adobe 软件产生的文档进行抽取，如 Photoshop 7.0,、Acrobat 5.0、 FrameMaker 7.0、GoLive 6.0、InDesign 2.0、InCopy 2.0、Illustrator 10.0、LiveMotion 2.0。

（2）可快速为已生成的元数据信息提供多种输出格式。Catalogue 对整个文件夹或多个文档进行识别，自动抽取出元数据信息，并可对自动生成元数据进行修改和补充。在抽取元数据前，用户可自定义需抽取元数据的字段。Catalogue 提供 HTML、CSV、Word、XML 格式的元数据报告，以后还可以生成 Execel 报告。XML 格式的元数据报告可直接用于数据交换和共享，还可以用 XML 专业工具将 XML 输出文档整合到元数据数据库中。通过对 XML 元数据文档进行适当的 XSL 转换，用户可生成 HTML、RTF、TXT、CSV 等格式的报告。

（3）可对一类 MS Office 和 Windows 2000 文档的属性进行修改。用户可以对整个文件夹（包括子文件夹）所包含的电子文档的属性值进行一致修改，例如改变作者、文件关键词属性等。

Catalogue 对电子文档元数据的抽取速度快，在很短时间内可完成对不同格式文档的批量抽取，并能以多种格式进行输出；一次能对最多 32767 篇电子文档进行元数据自动抽取，可满足海量文献描述的需要；软件简单易用，且抽取元数据字段较多。笔者用此软件的试用版本分别对中英文文献进行元数据抽取，结果显示其对英文文献的抽取效果较好，中文文献则不能实现元数据的正常抽取。

3.4 三种元数据自动抽取器功能比较

通过对以上三种典型的元数据自动抽取器的分析和试用，可将三者功能归纳为表 1：

表 1：三种元数据自动抽取器功能比较

	DROID	NLNZ Metadata Extractor	Metadata Miner Catalogue PRO
文档格式	PDF、DOC、RAR、JPG等546 种格式[15]	图象文件、办公文档、音频视频文件、标记语言文档四种类型，共 15 种格式	9 种格式：Microsoft Office、OpenOffice、StarOffice 、Visio documents 、HTML、PDF、JPEG、Tiff、PSD
抽取字段	PUID 、 MIME Type 、 Format 、 Version 、 Status、Warning	文档的名称、URL、类型、大小、时间、软硬件环境相关信息等	文档作者、大小、题名、关键词、时间、页数、行数、字符数等元数据信息
输出格式	XML、CSV	XML	HTML、CSV、Word、XML

抽取效果	抽取的元数据字段少；中英文文献都能识别	文档外部特征元数据多，内容特征很少涉及；对中文文献不能有效抽取	英文文献的内容特征元数据抽取效果好，中文文献抽取效果欠佳

4 元数据自动抽取技术的局限性

4.1 中文文献元数据自动抽取研究有待提高

国内外学者对电子文档元数据自动抽取的研究，大都关注对英文文献的抽取，中文文献的元数据抽取研究较少。相关论文都阐述了如何自动抽取英文文献的元数据信息，而对中文文献自动抽取的研究则相对薄弱。现有的元数据自动抽取器都是由国外组织和机构研制和开发的，对英文文献能进行有效抽取；而中文文献的抽取效果则不理想，有些文献不能完成元数据抽取，大部分中文文献的抽取结果存在着乱码。

4.2 电子文档的格式有所限定

目前元数据自动抽取的研究，基本上都支持对 PDF、DOC、PPT 等通用格式的文档进行自动抽取，对其他格式的电子文档的研究较少。数字图书馆建设和机构知识库中不仅存储着大量文本格式的电子文档，语音、图片、视频等多媒体信息也逐渐增多，所以对其他格式电子文档的元数据抽取研究有待发展和深入。

4.3 元数据自动抽取器对内容元数据的抽取效果欠佳

元数据自动抽取器对电子文档进行元数据自动抽取，基本上都是从电子文档的头文件中抽取，抽取的字段以文档的形式特征（类型、生成时间、软件相关信息等）为主，而关键的内容特征的相关元数据则难以获得。DROID 只能实现对电子文档的外部特征进行识别和抽取，其抽取字段相对较少；NLNZ Metadata Extractor 虽然提供了两种元数据生成方案，抽取的元数据字段也较多，但题名、作者、文摘、引文等元数据信息还不能实现有效抽取；Metadata Miner Catalogue PRO 在三种抽取器中表现最佳，不仅提供多种元数据生成格式，还实现了对英文文献作者、题名和部分文献关键词的自动抽取。三种元数据自动抽取器还不能完全满足现有的需求，对中文文献的抽取效果也有待提高。

4.4 未实现与数据库系统的有效集成

对元数据自动抽取技术的研究，没有做好与具体数据库系统的交互的研究。现有的相关研究中还没有将元数据自动抽取器与数据库系统集成的实例。元数据自动抽取的最终目的是为了便于使用，如能将自动抽取出来的元数据信息自动导入相关数据库系统或元数据仓储中，则能很好的解决元数据必须手工录入到数据库这一现实问题。

5 总结和展望

随着 e-Science、e-Research、e-Learning 等数字化科研环境、教学环境的迅速发展，如何有效组织和管理海量文献将成为图书情报机构的一大难题，这就迫切需要元数据自动生成技术的支撑。国内外对元数据抽取技术的研究和探索，为我们进行元数据自动抽取实践提供了一定的技术支持和理论依据。但现有的相关研究和实践，不能很好地满足数字图书馆和机构知识库建设中海量文献描述的需要。元数据抽取字段、文档语种、文档格式、文档类型等方面仍需改进，元数据自动抽取器抽取的元数据质量也不够高。总的来说，元

数据自动抽取技术还不够完善，仍有很多的问题亟待解决。笔者认为，将现有的元数据自动抽取器进行汉化，提高其处理中文文献的能力；并在此基础上进行二次开发，设计与具体数据库系统交互的开放接口，实现元数据自动生成并导入相关数据库，才能真正满足图书情报机构的需求。

（作者 E-mail:zengs@mail.las.ac.cn）

参考文献:

[1] Dublin Core metadata editor.[2007-11-8].http://www.ukoln.ac.uk/metadata/dcdot/

[2] Ying Liu, Kun Bai, Prasenjit Mitra, C. Lee Giles. TableSeer: Automatic Table Metadata Extraction and Searching in Digital Libraries[EB/OL]. [2007-11-10].
http://delivery.acm.org/10.1145/1260000/1255193/p91-liu.pdf?key1=1255193&key2=9007077911&coll=GUIDE&dl=GUIDE&CFID=9677192&CFTOKEN=66821516

[3] Min-Yuh Day , Richard Tzong-Han Tsai , et al. Reference metadata extraction using a hierarchical knowledge representation framework [J]. Decision Support Systems , 2007(43) : 152－167

[4] Eli Cortez1. Altigran S.da Silva1 , et al. FLUXCiM: Flexible Unsupervised Extraction of Citation Metadata[EB/OL]. [2007-12-18].
http://delivery.acm.org/10.1145/1260000/1255219/p215-cortez.pdf?key1=1255219&key2=9296088911&coll=GUIDE&dl=GUIDE&CFID=10613840&CFTOKEN=55320929

[5] Yunhua Hu, Hang Li, et al. Automatic extraction of titles from general documents using machine learning[J]. Information Processing and Management , 2006(42) : 1276－1293

[6] 贺亚锋. Web站点元数据自动生成工具介绍[J]. 图书馆杂志, 2001,20(1): 28-30

[7] Yewei Xue, Yunhua Hu. et al. Web page title extraction and its application[J]. Information Processing and Management . 2007 (43) : 1332－1347

[8] Jiangde Yu, Xiaozhong Fan . Metadata Extraction from Chinese Research Papers Based on Conditional Random Fields[EB/OL]. [2007-12-1].
http://210.37.44.253/nc2007/fskd2007/data/Volume%201/105-1-Chinese%20Research%20Papers.pdf

[9] 李朝光, 张铭, 邓志鸿, 杨冬青, 唐世渭. 论文元数据信息的自动抽取[J]. 计算机工程与应用 , 2002,38(21): 189-191,235

[10] DROID.[2007-11-22].http://droid.sourceforge.net/wiki/index.php/Introduction

[11] Metadata Extraction Tool.[2007-12-3].http://sourceforge.net/projects/meta-extractor/

[12] Nation Library of New Zealand .[2007-12-5].
http://www.natlib.govt.nz/about-us/current-initiatives/metadata-extraction-tool

[13] Catalogue PRO . [2007-12-8]. http://peccatte.karefil.com/software/Catalogue/catalogueDK.htm

[14] Main Features of Catalogue . [2007-12-10].
http://peccatte.karefil.com/software/Catalogue/CatalogueENG.htm

[15] Implementing the PREMIS data dictionary: a survey of approaches[EB/OL]. [2007-12-16].
http://www.loc.gov/standards/premis/implementation-report-woodyard.pdf