国家科学图书馆青年人才领域前沿项目研究报告（O80007）

元数据自动抽取工具在数字知识库建设中的应用研究与开发

中国科学院国家科学图书馆制

2009 年 11 月

元数据自动抽取工具在数字知识库建设中的应用研究与开发

项目总指导：马建霞

项目负责人：张秀秀

项目组成员：刘巍、卢利农、曾苏

目录

1 引言

1.1 研究背景及意义

随着计算机技术和网络技术的迅猛发展，e-Science、e-Research、e-Learning等数字化科研环境和教学环境的出现，科研机构和大学中的研究人员、教师、学生在研究过程中产生了大量的数字化文档。这些数字化文档主要包括已经发表或未发表的期刊论文、会议论文、专著、研究报告、学位论文、教学课件等，有必要对其进行长期保存和开放利用，因此数字知识库的建立就显得尤为重要。

在数字知识库中数字化资源是通过元数据来描述、组织和检索利用的，然而当前数字知识库的建设过程中，元数据大部分依赖作者、图书馆员逐条输入相关信息，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工作负担和极大地提高工作效率，进而加快数字知识库的建设步伐，使之更好地融入科研信息环境。

目前，国内外对元数据自动抽取技术已有许多研究，其实现方法大体上可以分为两类，即基于规则的方法和机器学习的方法，它们各有优缺点。基于规则的方法采用基于模式识别和模式匹配的模版挖掘技术达到抽取自由文本的目的，其优点是易于理解和操作，并且如果抽取规则制定得当，抽取效果将十分理想。但是基于规则的方法需要专业人员预先设计一系列规则，而且如果抽取的目标发生变化则会有规则不适应的情况出现。机器学习的方法采用另外一种思路，它通过训练样本并建立样本的输入与输出之间的关系来预测新数据，其优点是该方法具有良好的适应性，但建立的模型的有效性依赖于训练样本的数量和质量。

另外，互联网中也出现了许多数据抽取工具，例如由 Sytec Resources 为新西兰国家图书馆开发的 NLNZ Metadata Extractor 和法国 Soft Experience 开发的 Metadata Miner Catalogue PRO 就是专门用于处理数字化文档和提取元数据信息的工具。此外，SOFTPEDIA 网站也发布了近 40 款与数据抽取相关的软件。然而直到目前可以说这些软件的功能还不够完善，它们绝大多数关心的是文档属性元数据（如创建日期、文件大小、资源标识等）的抽取，仅有少部分工具支持文

档内容元数据（如题名、作者、摘要等）的抽取，并且抽取中文文档的效果也不甚理想。

1.2 主要研究工作

本项目旨在研究国内外的相关应用，并在分析几种典型文档格式（PDF、DOC、PPT）的基础上，结合现有开源软件的应用和功能改进，探索从特定文档格式的文本中自动抽取有关内容的元数据，并应用到数字知识库建设中的可行的技术路线，并在基于 DSpace 的机构知识库建设中进行实验性应用和验证。

本项目将利用 Java 语言完成元数据自动抽取工具开发，并提供调用接口方便与数字知识库系统的集成。另外，元数据自动抽取工具能够在主流系统平台运行，例如 Windows/Linux/Unix。

1.3 研究方法及路线

本项目将运用文献调研法搜集国内外相关技术资料，运用对比分析法对开源元数据抽取工具进行使用测试及对比分析，最后运用系统原型法实现元数据自动抽取工具的功能。

项目的实施路线为：文档格式（特点）分析→现有工具（技术）分析→模块与算法设计→编码与调试→系统测试与应用，其过程可以用图 1-1 表示。

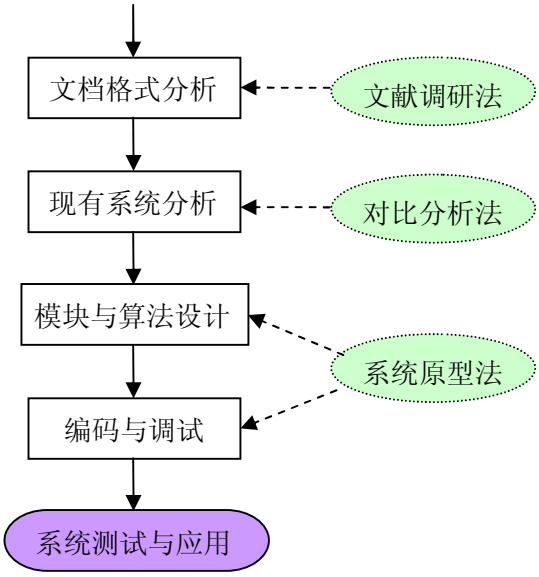

图 1-1 项目的实施路线图

2 元数据自动抽取的国内外研究现状

2.1 元数据自动抽取的相关研究

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：

2.1.1 对特定格式文档的元数据自动抽取

数字化文档的类型主要有 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。

2.1.2 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[3]等人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。文献[4]抽取的也是引文元数据，其抽取结果构建了一个本体模型，为基于语义的检索提供了基础。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。

---

[1] Dublin Core metadata editor. 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].

[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] 郭志鑫. 基于本体的文档引文元数据信息抽取[J]. 微计算机信息, 2006(22): 304-306

[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

2.1.3 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 计划的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。

2.1.4 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、其他类型的文档进行元数据自动抽取则不能完成。

2.2 元数据自动抽取工具

元数据抽取是信息抽取的一个分支，因此在互联网中出现的众多数据抽取（data extractor）工具并不能完全满足描述性元数据自动抽取的需要[10]。目前发现的元数据自动抽取工具主要有：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 NLNZ Metadata Extractor 软件和法国的 Metadata Miner Catalogue PRO 软件。DROID 和 NLNZ 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，其免费提供的试用版仅能提供十条数据的批处理能力。

---

[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].

[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,(21)

[10] SOFTPEDIA. http://www.softpedia.com/

2.2.1 DROID

DROID[11]（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 V4.0）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。图 2-1 展示了 DROID 的工作快照。

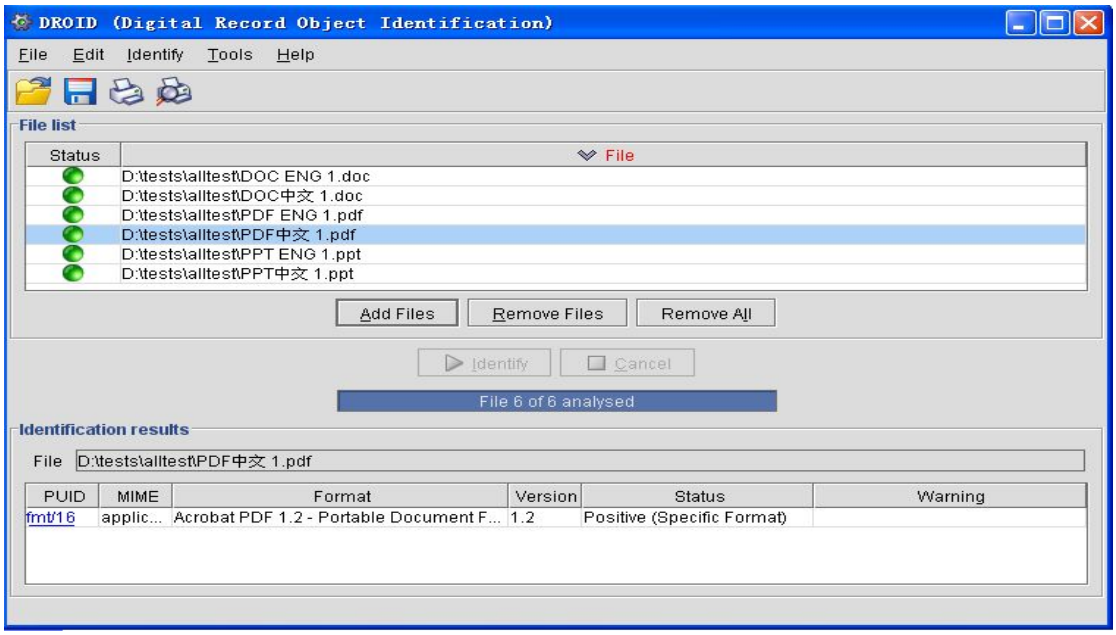

图 2-1 DROID 的工作快照

目前，DROID 仅能实现对文件 PUID（PRONOM 唯一标识符）、MIME Type

[11] DROID. http://sourceforge.net/projects/droid/files/

（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 只能对数字化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。此外，DROID 仅支持 PDF、DOC、RAR、JPG 等常用格式文档的识别，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。

2.2.2 NLNZ Metadata Extractor

NLNZ Metadata Extractor[12]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 V3.4 版本，从 2007 年开始可免费下载使用。NLNZ 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。

NLNZ Metadata Extractor 采用“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）。

NLNZ Metadata Extractor 对硬件环境要求低，在一般的 PC 上都可运行，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；抽取的结果不是

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

显示在工作界面上，而是以 XML 形式保存到文件中。但该抽取器从电子文档的头文件中抽取元数据信息，不能对电子文档的内容进行抽取。抽取的字段大都是文档的属性元数据，如文档名称、类型、修改时间、URL、软件版本等，内容元数据如题名、作者、文摘、引文等字段均没有反映；而且对中文文档中出现的中文字符无法识别，只能显示出英文和数字部分。图 2-2 展示了 NLNZ Metadata Extractor 的工作快照，图 2-3 展示了 NLNZ Metadata Extractor 抽取结果（XML 文件输出）。

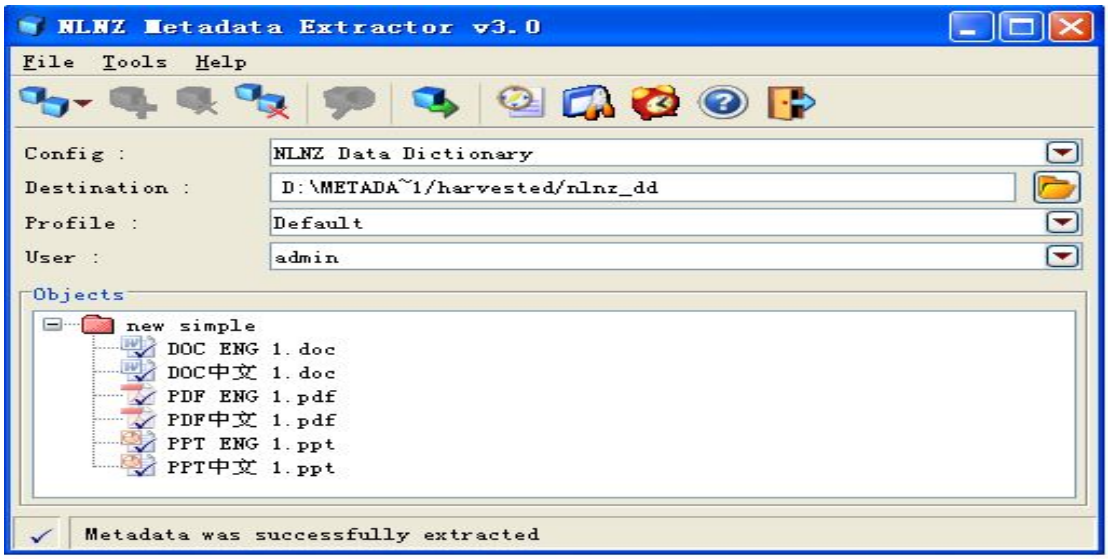

图 2-2 NLNZ Metadata Extractor 的工作快照

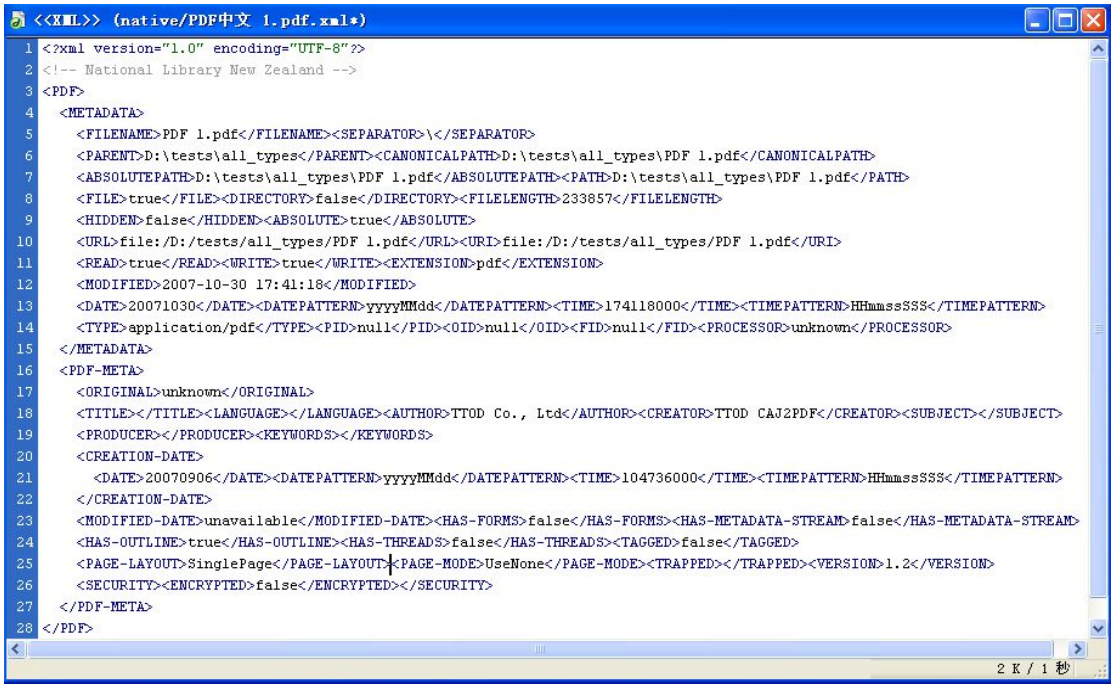

图 2-3 NLNZ Metadata Extractor 的抽取结果

2.2.3 Metadata Miner Catalogue PRO

Metadata Miner Catalogue PRO[13]（以下简称 Catalogue）是由 Soft Experience 开发的商业软件，可用于题名、作者、主题、关键词等描述性元数据的自动抽取。Catalogue 能够处理的文档类型有 Microsoft Office、OpenOffice、StarOffice、Visio documents、HTML、PDF、JPEG、Tiff、PSD 等多种，并且提供英语、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：

（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，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；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 格式图像文件，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）格式文件，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 Excel 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（含子文件夹）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

[13] Metadata Miner Catalogue PRO. http://peccatte.karefil.com/software/Catalogue/MetadataMiner.htm

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（付费版）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。项目组用此软件的免费试用版分别对中英文文献进行元数据抽取，结果显示该软件对文件名称的抽取效果较好，而对其他的内容元数据（如作者、关键词、摘要等）抽取效果较差。

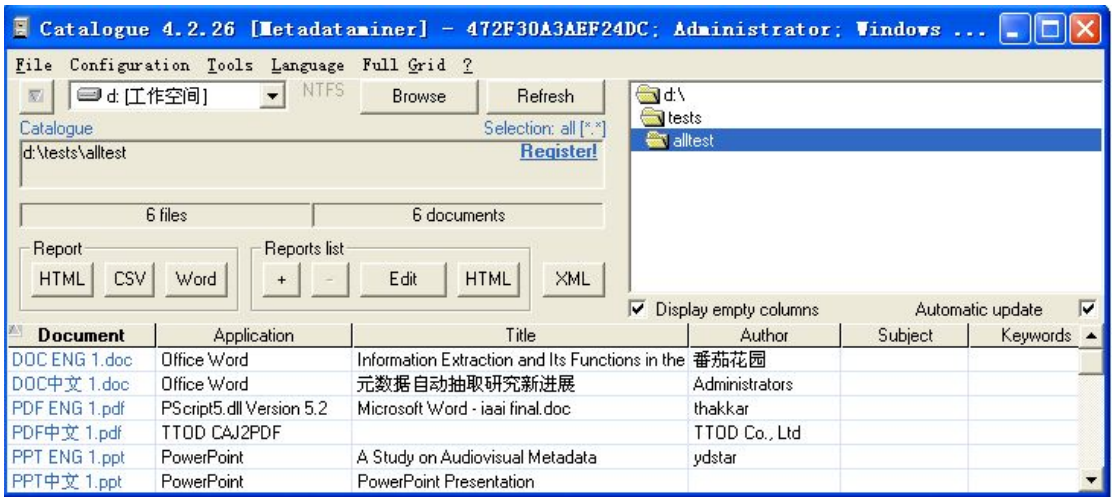

图 2-4 Catalogue 的工作快照

2.3 元数据自动抽取的局限性

2.3.1 中文文档元数据自动抽取有待提高

目前可获得的元数据自动抽取工具都是由国外组织和机构研制开发的，对英文文档能进行有效处理，而中文文档的抽取效果则不甚理想，有些连中文字符都无法识别，大部分中文文献的抽取结果都不够准确。

2.3.2 自动抽取的元数据质量不高

目前的元数据自动抽取工具多数都是从电子文档的头部抽取信息，抽取的字段以文档的属性特征（类型、生成时间、软件相关信息等）为主，而描述一个文件的关键的内容特征则难以获得。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，还实现了对英文文档作者、题名和部分文档中关键词的自动抽取。三种元数据自动抽取器还不能完全满足现有的需求。

2.3.3 未实现与数字知识库的有效集成

对元数据自动抽取的研究，没有做到与具体数字知识库系统的有效集成。元数据自动抽取的最终目的是为了便于使用，如能将元数据自动抽取应用到数字知识库或元数据仓储的流程中，则能很好的解决元数据必须手工录入这一现实问题。

3 元数据自动抽取工具

3.1 系统设计

元数据自动抽取工具的主体由四个模块组成，分别是格式识别器、匹配器、XML 模版和接口适配器，其系统结构如图 3-1 所示。格式识别器通过 Java 语言 File 类提供的 file.getName().toLowerCase().endsWith()方法判断输入的源文件格式，然后将其转入不同的匹配器进行元数据自动抽取（工具仅对 PDF、DOC 和 PPT 三种类型文档进行处理，其他类型的文档忽略不处理）；匹配器分为三种，分别对应 PDF、DOC 和 PPT 类型文档进行文件结构和文本内容解析，并根据预定义的一系列抽取规则完成描述性元数据（包括属性元数据和内容元数据）的自动抽取，是元数据自动抽取工具的核心。其中 PDF 匹配器是在 PDFBox 开源软件包的基础上开发，而 DOC 匹配器和 PPT 匹配器则是在 Apache POI 开源软件包的基础上开发。XML 模版采用 DC 元数据标准，由匹配器自动填充数据并输出；接口适配器根据特定数字知识库的元数据需要实现定制和映射。

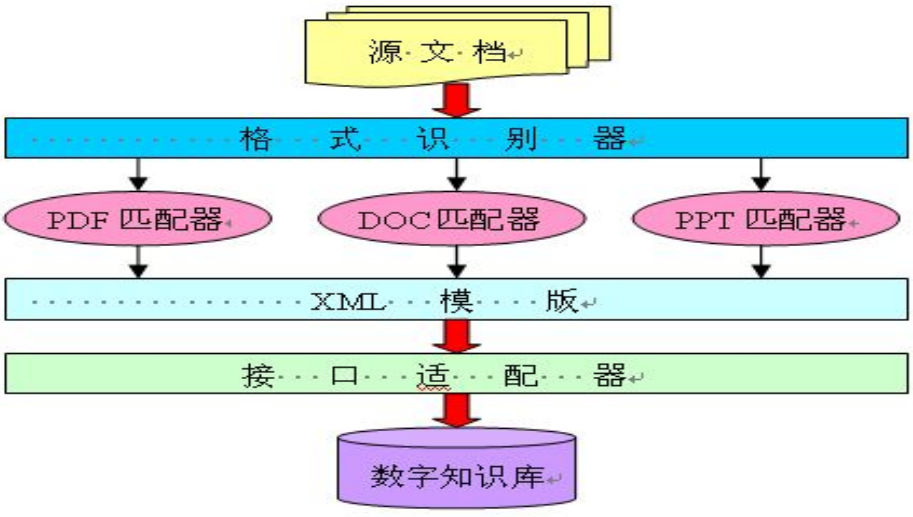

图 3-1 元数据自动抽取工具的系统结构图

3.2 抽取规则

3.2.1 内容元数据抽取规则的制定原则

在前期资料搜集过程中，由于获得了较多关于 PDF 文档的技术资料，其中 Adobe 公司官方网站上公布的 PDF Reference[14]更是关于 PDF 文档的权威详细解析，考虑到数字知识库中 PDF 类型文档里科技论文占了很大一部分，所以 PDF 匹配器中关于PDF文件的内容元数据抽取规则主要从科技论文的角度参考制定。而微软 Office 办公套件由于商业保密所限，关于 DOC 和 PPT 文档结构的特点也仅获得了初级了解[15]，所以 DOC 匹配器和 PPT 匹配器中关于 DOC 和 PPT 的内容元数据抽取规则制定的较为简单。

3.2.2 PDF 文档内容元数据抽取规则

科技论文是自由格式的文本组合，不同的出版商在论文排版方面有着不同的规定，这就决定了内容元数据的自动抽取具有一定的难度。但论文信息的组织仍有一定的规律可寻，经研究发现，大部分论文的框架都可以分为如下 6 个部分：

① 标题（可以有副标题）；

② 作者及相关信息（可以有多个）；

③ 摘要；

④ 关键词（可以没有，英文文章不太注重关键字）；

⑤ 论文主体；

⑥ 参考文献。

从描述一篇论文的角度看，主要关心前 4 个部分，因为它们基本涵盖了整篇论文的主要内容。另外，前 4 个部分基本上都出现在论文的第一页，所以为了提高抽取效率，在实际处理过程中，仅对 PDF 文件的第一页进行了结构和内容的解析。

（1）标题的抽取

标题一般没有什么固定的位置，比如有些文章可能包含页眉信息，此时标题

[14] PDF Reference. http://www.adobe.com/devnet/pdf/pdfs/PDFReference13.pdf

[15] 李秀芹，朱跃龙. Microsoft Word 文件转换器的设计与实现[J]. 计算机应用, 2002(22): 37-38,41

会出现在页眉以下；有些文章可能没有页眉信息，此时标题会出现在文章的第一行。另外，科技论文的研究领域涉及方方面面，因此标题也没有一个专用名词供识别。不过，绝大多数论文标题的字体都是整篇文章中最大的，因此可以根据标题的这一特征来定位和抽取。

PDF是一种标签命令式的结构化文档格式，在PDF的众多标签命令中，以BT操作符开始、以ET操作符结束就标识了一个文本对象。文本对象中既包括了文本内容本身（Tj/TJ操作符用来设置文本，括号内的参数就是希望获得的文本串），也包括了显示所依赖的字体（Tf 操作符用来设置字体，它的第一个参数描述字体名称，第二个参数描述字体大小。第二个参数的值越大，说明字体越大，反之则越小。另外，英文PDF文件习惯将Tf 的第二个参数值设为1.0，此时可以从Tm操作符获得字体信息。Tm操作符共有6个参数，其中第一个参数基本反映了字体大小）、位置（PDF文件将打印区的左下角设置为打印原点，y轴正方向朝上，x轴正方向朝右。Td/TD操作符可以设置文本行的位置，第一个参数描述当前行的水平位移，第二个参数描述当前行的垂直位移）等格式信息。

图3-2是某中文科技论文的文件头信息，图3-3是其文件在解码后包含了两个文本对象的部分内容流。

图3-2 一个PDF格式的中文科技论文的例子

```
BT
/F0 16.0 Tf
47.8 613.0 Td
[(图)17.5(书)17.5(馆)17.5(自)17.5(动)17.5(化)] TJ
/F0 21.0 Tf
110.2 564.9 Td
[(上)1.9(海)1.9(图)1.9(书)1.9(馆)1.9(新)1.9(馆)1.9(计)1.9(算)1.9(机)1.9(管)1.9(理)1.9(系)1.9(统)] TJ
ET

BT
/F0 14.0 Tf
235.7 535.4 Td
[(葛)1.9(如)1.9(琛)] TJ
ET
```

图3-3 解码后包含了两个文本对象的部分内容流

有些文章可能会有副标题，副标题的字体一般都比标题小，而且位于标题以下，另外对于中文文章，副标题一般会以破折号“——”开始。

（2）作者名的抽取

作者名的抽取工作最为复杂，因为不同文献处理作者及相关信息的排版方式种类最为繁多，而且中英文文献略有差异。总体来说，作者名通常位于标题的下方、地址或邮件等的上方，可能会有一个或多个作者，但大多会在一行排列。中文文章伴随作者名的通常有作者单位信息，放在一对圆括号中，而英文文章伴随作者名的有作者单位信息，或者还有 E_mail 信息。因此，在具体实现中，首先定位标题，如果标题以后不是副标题，那么就可以抽取作者信息了。但是怎样判断抽取结束呢？可以考虑下面几种情况：

① 下一行是否以左括号开始；

② 下一行中是否含有标识作者单位的名词，如 Department、Center、School、University、Institute；

③ 下一行中是否含有标识作者 E_mail 的文本符号“@”；

④ 下一行是否遇到标识摘要的专用名词“摘要”或者“Abstract”。

如果遇到上述四种情况中的任何一种，都标志着作者名抽取结束。

（3）摘要的抽取

不论是中文摘要，还是英文摘要，通常都有一个专用名词供识别，即：

“摘要”+摘要描述，

或者

“Abstract”+Description。

一旦匹配到上述规则的表达式，就可以获取摘要信息了。

（4）关键词的抽取

关键词也有一个专用名词供识别，即：

“关键词”+关键词表，

或者

“Keywords”+keyword list。

一旦匹配到了上述规则的表达式，则可以获取关键词信息了。

3.2.3 DOC 文档内容元数据抽取规则

考虑到数字知识库中 DOC 文档多为科技论文在正式发表前的预印本，所以关于 DOC 文档的内容元数据抽取规则与 PDF 文档的类似，但是由于无法获知字体、位置等具体的格式信息，所以关于标题的抽取规则要简单许多。在 DOC 文档中，认为第一行文本行即为该文档的标题。其余关于副标题、作者、摘要、关键词的抽取规则与 PDF 文档完全一致。

3.2.4 PPT 文档内容元数据抽取规则

考虑到通常的 PPT 文档中没有摘要和关键词，所以对 PPT 文档的内容元数据抽取仅做了标题和作者两项。对于标题认为第一行文本行即为该文档的标题，而紧接着的第二行则认为是作者/机构信息。

3.3 实验测试

元数据自动抽取工具能够对 PDF、DOC、PPT 三种格式的数字化文档进行批量处理，其抽取的元数据信息包括：文档属性元数据，包括创建日期、文件大小、资源标识等；文档内容元数据，包括题名、作者、关键词、摘要。另外，元数据自动抽取工具能够以 XML 格式保存抽取结果，并将其存放到用户指定的目录下，而且该工具能够跨平台应用，如 windows、Linux。

图 3-4 是元数据自动抽取工具在 Windows 系统中的工作快照，图 3-5 是其 XML 格式的输出结果，图 3-6 是其在 Linux 环境下的工作快照。

图 3-4 元数据自动抽取工具在 Windows 系统中的工作快照

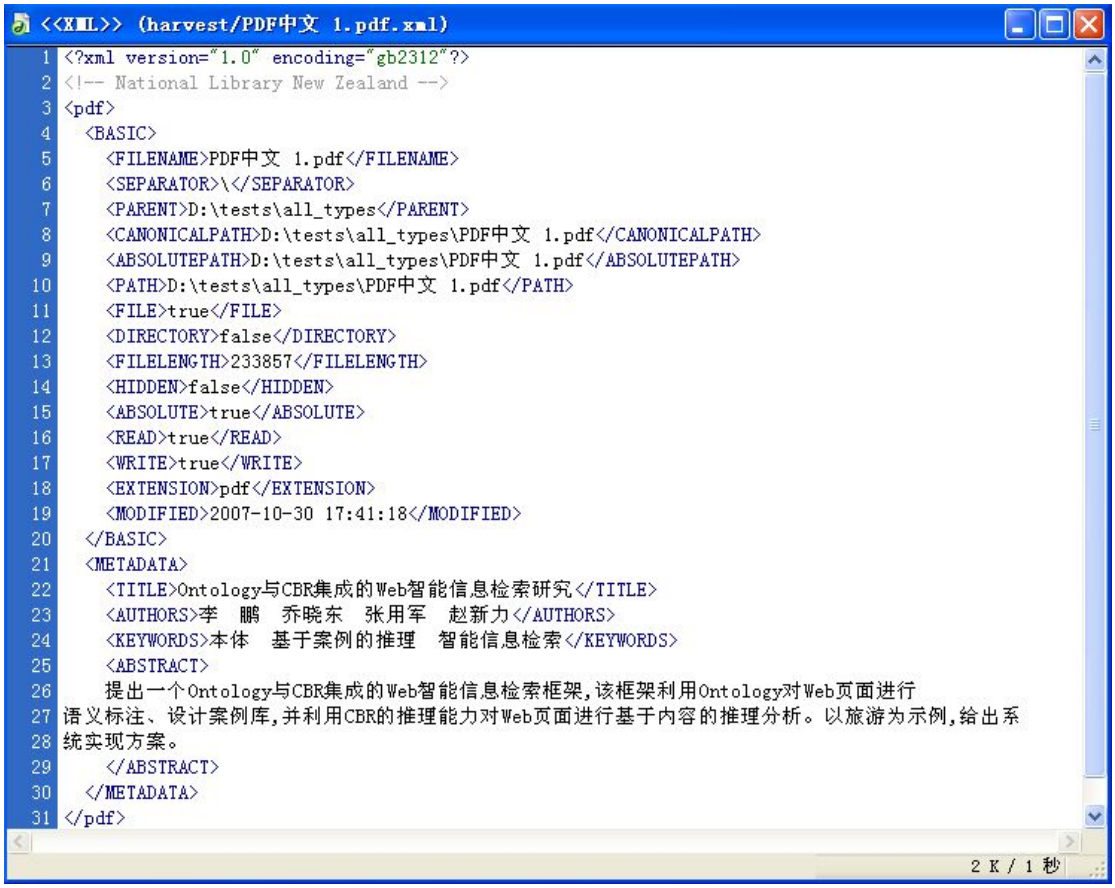

图 3-5 元数据自动抽取工具的 XML 输出结果

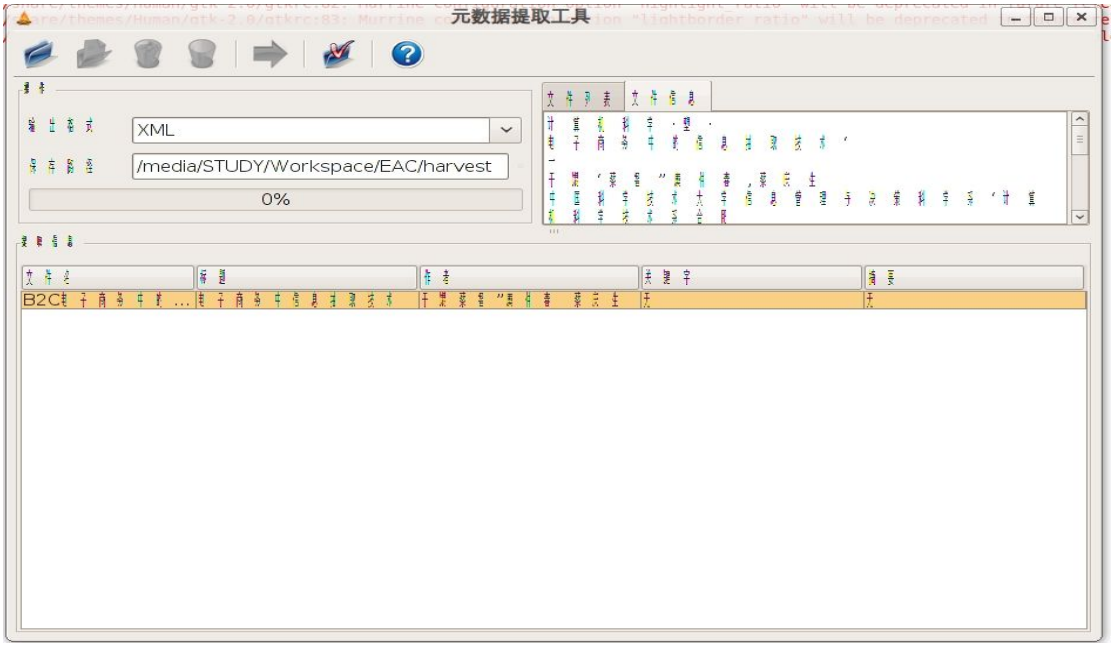

图 3-6 元数据自动抽取工具在 Linux 环境下的工作快照

为了评价元数据自动抽取工具的抽取效果，我们对PDF类型的中英文科技论文分别进行了准确率测试。其中中文测试集来源于《中国学术期刊全文数据库》，以“信息抽取”为关键词进行精确检索，共检索到文献213篇（2008年10月31日），实际下载199篇；英文测试集来源于Springer，以“metadata”为关键词进行检索，共检索到文献11426篇（2008年11月2日），实际下载了前200篇。测试结果见表1。

表1 元数据自动抽取的实验结果

	中文	英文
标题	0.841	0.850
作者名	0.708	0.683
摘要	0.914	0.930
关键词	0.901	0.974

从表1可以看出，元数据自动抽取工具基本上能够较好地完成PDF科技论文的语义元数据抽取。但是由于不同的期刊具有不同的论文版式，即便是同一种期刊，不同主题的文献其版式也会有一定的差别，这就使得抽取结果不可避免的出现一定程度的偏差。

总体上，摘要和关键词抽取的准确率较高，而中英文标题抽取的准确率分别为84.1%和85.0%。造成标题无法正确抽取的原因可能有：

A. 标题并不是论文首页中字体最大的，此类文献多见于《中文信息学报》；

B. 某些未知原因使得解析的文本中有部分文字显示为乱码；

C. 论文可能是以扫描方式上传的，因此解析的内容流中提取不到文本信息。

中英文作者名抽取的准确率最低，分别为70.8%和68.3%。影响作者名抽取的准确率的原因可能有：

A．标题定位错误，造成作者名的抽取规则失效；

B．某些未知原因使得解析的文本中有部分文字显示为乱码；

C、论文可能是以扫描方式上传的，因此解析的内容流中提取不到文本信息；

D．作者名有规则以外的排版方式没有定义，如作者名出现在标题前等。

4 元数据自动抽取在 Dspace 实验系统中应用

本项目建立的 Dspace 实验系统是在中国科学院国家科学图书馆机构知识库的基础上修改完善的。在 Dspace 实验系统中数字化文档的提交流程增加了元数据自动抽取功能，用户进入“我的工作间”开始一个新的提交，首先上传文件并由系统自动抽取元数据，之后用户核实自动抽取的元数据信息是否有误并补充其它必要元数据后结束提交流程，整个过程如图 4-1 和图 4-2 所示。

图 4-1 DSpace 试验系统的提交流程：上传文件（以 PPT 文档为例）

图 4-2 DSpace 试验系统的提交流程：填充元数据（以 PPT 文档为例）