PDF 科技论文语义元数据的自动抽取研究*

张秀秀　　马建霞
（中国科学院国家科学图书馆兰州分馆 兰州 730000）

[摘要] 分析了 PDF 文件结构，在此基础上，解析 PDF 文件的内容流，并采用基于规则的匹配方法和基于格式的定位方法，自动抽取科技论文中的语义元数据信息。实验结果表明，上述方法对标题、作者等重要的语义元数据信息实现了较好的抽取效果。

[关键词] PDF　科技论文　语义元数据　自动抽取
[分类号] TP391.43

Automatic Extraction of Semantic Metadata from PDF Research Papers

Zhang Xiuxiu　　Ma Jianxia
(Lanzhou Branch of National Science Library, Chinese Academy of Sciences, Lanzhou 730000, China)

[Abstract] In order to extract semantic metadata from PDF research papers, we analyze the structure of PDF documents, On the basis of that we parse the content streams and give a hybrid method combining rule-based matching approach with format-based locating approach. Experiments show that our extractor gains a fairly high effect.

[Keywords] PDF; Research paper; Semantic metadata; Automatic extraction

1　引言

元数据提供了网络资源描述、表达、管理和使用的基本方案，是网络资源组织和检索的核心所在。随着计算机技术和网络技术的迅猛发展，面对海量文献描述的需要，如何快速、高效地产生元数据成为数字图书馆建设过程中面临的一大难题。当前数字图书馆建设过程中，元数据大多由人工逐条标记输入，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管理的需要。若元数据信息可以自动生成、自动抽取，必将大大减轻信息人员的工作负担，极大地提高工作效率。

目前，网上发布的科技论文大多以 PDF 形式存在，因此，本文的研究将针对 PDF 格式的论文展开。文章首先介绍了 PDF 文件的物理结构和逻辑结构，然后在对 PDF 文件直接进行文本、格式解析的基础上，依据科技论文中文本内容的组织方式和排版格式等信息，采用基于规则的匹配方法和基于格式的定位方法，实现相关元数据的自动抽取，其中最主要的工作是抽取出论文的标题、作者、摘要、关键词四种重要的语义元数据。

2　元数据自动抽取的相关研究

元数据抽取是信息抽取的一个分支，随着元数据自动抽取的内在需求不断增长，国内外学者对元数据自动抽取技术展开了一系列的理论研究。

目前，元数据自动抽取的方法大体可以分为两类，基于规则的方法和机器学习的方法。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，如文献[1]利用正则表达式规则从 PDF 文档中抽取首页元数据；文献[2]采用基于层级知识描述框架的 InfoMap 方法抽取引文元数据等。基

* 本文系中国科学院国家科学图书馆青年人才领域前沿项目“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”及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。

于规则的方法易于理解和操作，并且如果规则制定得当，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，如文献[3]采用最大熵等模型从常见文档中抽取标题元数据；文献[4]采用条件随机场模型抽取多种通用元数据；文献[5]采用概率评估模型抽取引文元数据等。机器学习的方法具有良好的适应性，但机器学习的方法建立起来的模型，其有效性依赖于训练样本的数量和质量。

另外，文献[6]利用PDF2HTML工具将PDF格式的文件转化成XML格式的中间文档，再利用转化过程中保留的文件格式信息抽取论文的首页元数据。利用文件格式信息抽取元数据启发了元数据自动抽取的新思路，改变了基于规则的方法和机器学习的方法只能从结构松散或者纯文本中抽取有用信息的一贯做法，但是目前的PDF转化工具转化效果参差不齐，难免在转化过程中造成一些格式信息的失真。

本文的工作主要借鉴了文献[1]和文献[6]的研究成果，不同之处在于文献[1]中元数据抽取仅仅依赖基于规则的方法，处理的对象是自由文本，而本文对于具有明显文本特征的关键词和摘要采用基于规则的方法抽取，对于论文标题和作者则更多地利用格式特征进行定位。另外，本文与文献[6]的抽取工作也有差别，文献[6]利用工具 PDF2HTML 首先将 PDF 文件转化，实际处理的对象是转化后的 XML 文档，而本文将解析 PDF 文件并直接获取文件中的文本与格式信息。

3 PDF 文件结构

PDF是一种标签命令式的结构化文档格式，支持7位ASCII码和多种压缩编码方式。一个原始的PDF文件从结构上可以分为四个部分：文件头、文件体、交叉引用表和文件尾[7]。文件头（Header）指明了文件所遵从的PDF规范的版本号，它出现在PDF文件的第一行。文件体（Body）是PDF文件的主体部分，由许多序列化的间接对象组成，这些间接对象共同构成了PDF文件的具体内容，如页面、字体、图像等。交叉引用表（Cross-reference Table）是一个关于间接对象的地址索引表，通过它能够实现对间接对象的快速随机存取。文件尾（Trailer）声明了交叉引用表的地址，指明了文件体的根对象，还保存了加密等安全信息。

PDF的文档结构反映了文件体中间接对象之间的等级层次关系，是一种树型结构，如图1所示。树的根节点就是整个PDF文件的根对象(Catalog)，根对象包含多种属性，其中最重要的属性为页面属性，它包含了PDF文件用于显示文字、图形、图像等的内容信息。

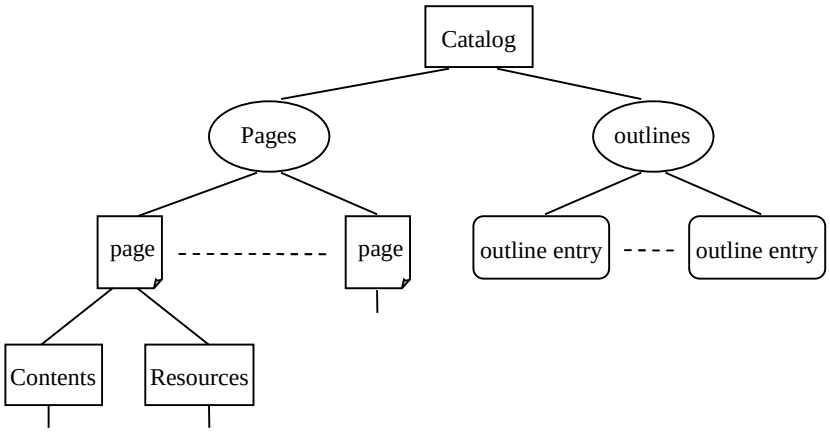

图1 PDF的文档结构

4 语义元数据自动抽取的设计实现

4.1 PDF 文件解析

根据文件尾提供的信息，可以找到交叉引用表和整个文件的根对象，从而读取 PDF 文件。所以整个处理流程将从寻找文件尾的 Trailer 关键字开始，具体步骤如下：

①从文件尾中找到属性标签/Root，取得其后的间接对象号，这个对象号标识了文档根对象的位置，是整个正文内容的入口；

②转入文档根对象，其标识为/Type /Catalog，在其中寻找属性标签/Pages，取得其后的间接对象号，这个对象号标识了文档页根对象的位置；

③转入文档页根对象，其标识为/Type /Pages，在其中寻找属性标签/Kids，取得其后的第一个间接对象号，这个对象号或者标识了文件第一页的对象位置，或者仍然是页根对象。如果情况为后者，则仍然执行步骤③，否则执行步骤④；

④转入文档页对象，其标识为/Type /Page，在其中寻找属性标签/Contents，如果找不到 Contents 标签，则说明此页内容为空，否则取得其后的全部间接对象号，并以先序深度优先的顺序按步骤⑤依次处理这些内容对象；

⑤转入内容对象，提取/Filter标签后的解码名，并将stream与endstream之间的内容流存入一个字节数组中。在源代码中，字节数组显示的内容流为乱码，需要对其进行解码处理。

PDFBox开源软件包中的filer包提供了有关解码的方法，根据解码名调用相应的解码方法，可以获得解码后的内容流。

⑥将所有内容对象的解码流连接起来，组成第一页的内容流。图2显示了某中文科技论文的文件头信息，图3显示了其解码后的部分内容流。

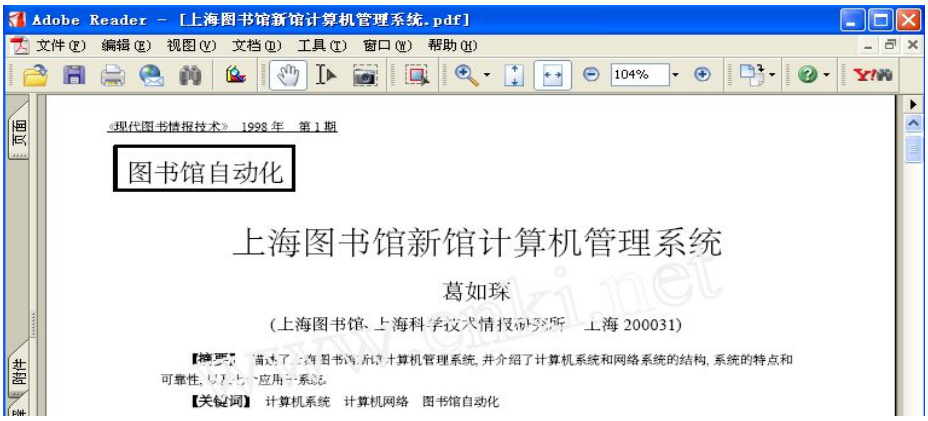

图2 一个PDF格式的科技论文的例子

```
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 解码后的部分内容流

文本对象：以BT操作符开始，以ET操作符结束，其内容既包括文本信息，也包括字体、位置等格式信息；

字体信息：Tf操作符用来设置字体信息，第一个参数描述字体名称，第二个参数描述字体大小，值越大，说明字体越大，反之则越小。另外，英文的PDF文件习惯将Tf的第二个参数值设为1.0，此时要从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上反映了字体大小。

位置信息：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；

文本信息：Tj/TJ操作符用来设置文本内容，括号内的参数就是希望获得的文本串。

4.2 内容元数据抽取分析

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 文章主体；
⑥ 参考文献。

从抽取的角度看，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，绝大多数文章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。

具体实现中，通过扫描整个内容字符串，寻找所有Tf操作符并获得第二个参数的值，比较得出最大者。如果所有Tf操作符的第二个参数值均为1.0，此时寻找所有Tm操作符并比较得出第一个参数值中的最大者。对应Td/TD操作符位置上的文本串就是标题。

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。

5 试验结果

图4为项目组成员基于Java语言自主开发的元数据自动抽取工具。该工具能够自动批量地抽取中英文的科技论文，并且在图形界面上显示标题、作者、关键词和摘要四种重要的语义元数据信息。

图4 元数据自动抽取工具显示界面

为了评价该工具的抽取效果，我们对中英文的科技论文分别进行准确率测试。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；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，实际下载了前200篇。实验结果见表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，即便同一种期刊，不同类型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；

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。

6 结论

采用基于规则的匹配方法和基于格式的定位方法，可以解决大部分 PDF 格式科技论文的语义元数据抽取。但是，毕竟没有任何规则可以涵盖现实世界中的所有情况，总会有规则之外的情况出现，使得元数据抽取的准确率降低。因而，对

于首页元数据的格式特征和文本特征等方面的总结还需要进一步完善。

参考文献:

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

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

[3] Yunhua Hu, Hang Li, Yunbo Cao, et al. Automatic extraction of titles from general documents using machine learning[J]. *Information Processing and Management*, 2006,42(1):1276 - 1293

[4] Jiangde Yu, Xiaozhong Fan. Metadata Extraction from Chinese Research Papers Based on Conditional Random Fields[J/OL]. [2008-10-21] http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4405975&isnumber=4405869

[5] C.L.Giles, K.D.Bollacker, S.Lawrence. CiteSeer: an automatic citation indexing system[J/OL]. [2008-10-21] http://clgiles.ist.psu.edu/papers/DL-1998-citeseer.pdf

[6] 陈俊林, 张文德. 基于XSLT的PDF论文元数据的优化抽取. *现代图书情报技术*, 2007(2):18-23.

[7] PDF Reference[EB/OL]. [2008-4-15]. http://www.adobe.com/devnet/pdf/pdfs/PDFReference13.pdf

(作者 E-mail:zhangxx@llas.ac.cn)

[作者简介] 张秀秀，女，1981年生，馆员；
马建霞，女，1972年生，副研究馆员。