基于 PDFBox 抽取学术论文信息的实现

牛永洁，薛苏琴
（延安大学 数学与计算机学院，陕西 延安 716000）

摘 要：为了对学术动态、热点及学术发展趋势进行研究，需要对学术研究论文进行数据挖掘研究。首先需要从海量的学术论文中提取有兴趣的信息。针对目前学术论文大多采用 PDF 格式的现状，重点研究了 PDF 文件的格式以及对 PDF 格式操作的各种技术，采用开源函数库 PDFBox 对 PDF 格式的学术论文按照规则进行信息的提取，提取的信息主要包括学术论文的标题、作者、单位、关键词、发表时间、摘要等信息。最后对提取信息的正确率进行了统计，有助于针对学术研究的大数据研究。

关键词：数据挖掘；信息抽取；PDF 格式；学术论文
中图分类号：TP39 **文献标识码**：A **文章编号**：1673 - 629X(2014) 12 - 0061 - 03
doi：10.3969/j.issn.1673 - 629X.2014.12.015

Realization of Extraction of Academic Papers Information Based on PDFBox

NIU Yong - jie，XUE Su - qin
（College of Mathematics & Computer，Yan'an University，Yan'an 716000，China）

Abstract：In order to research the academic dynamics，hot topic and academic development trends，need to carry out the data mining research for academic research papers. First of all，extract interest information from the massive papers. For the situation that the current academic papers are mostly used PDF format，mainly study the format of PDF files and a variety of technical operations for PDF operations，open - source library PDFBox is used to extract information for the academic papers with PDF format in accordance with the rules，the extracted information is mainly including academic titles，authors，unit，keyword，publication time，abstract and other information. Finally，the correct rate of extraction of information has been statistical，which is helpful for big data for academic research.

Key words：data mining；information extraction；PDF format；academic papers

0 引 言

随着互联网和通信技术的发展，大数据时代已经悄然来临。面对海量的信息与数据资源，人们常常面临数据量大，信息匮乏，因此对如何能够从中获取其隐含的、潜在有用的知识的要求变得很迫切，于是数据挖掘应运而生。数据挖掘的第一步就是数据的采集，能够快速、准确地采集到感兴趣的信息是数据挖掘的重要基础。

针对日益增多的海量的学术研究论文，对其进行数据挖掘可以发现某个学科的学术划分的细化程度，而且能够掌握学科的技术发展动态和学科发展趋势及发展速度，更能够掌握学术发展的热点区域。因此对学术研究论文进行数据挖掘是非常有意义的。

数据挖掘包含数据采集，数据清洗、数据存储、数据传输、数据分析、数据解释等过程，其中数据采集是数据挖掘过程的开始，是整个数据挖掘过程的基础和必要步骤。为了对学术研究论文进行数据挖掘，首先必须从论文中提取感兴趣的数据，目前国内外学术论文大多以 PDF 格式的形式存在，而数据挖掘的后续步骤直接对 PDF 格式的数据进行处理比较困难，因此需要从 PDF 格式的文件中将感兴趣的数据提取出来。

1 PDF 格式分析

结构化的文档格式 PDF（Portable Document Format）是由美国排版与图像处理软件公司 Adobe 于 1993 年首次提出的，已经有过 7 个版本，6 次版本升

收稿日期：2014 - 02 - 20 **修回日期**：2014 - 05 - 25 **网络出版时间**：2014 - 10 - 23
基金项目：陕西省自然科学基础研究计划项目（2013JM8042）
作者简介：牛永洁（1977 - ），男，河南许昌人，讲师，硕士，CCF 会员，研究方向为软件工程、数据挖掘。
网络出版地址：http://www.cnki.net/kcms/detail/61.1450.TP.20141023.1124.034.html

级,从最初的 PDF1.0.6 版本到现在的 PDF1.6。PDF 格式和已经熟知的 HTML、XML 等结构化的文件格式一样,包含有关键字、分隔符、数据等,不同的是 PDF 文件是按照二进制流的方式保存的,而 HTML 文件则是以文本方式保存。

PDF 文件物理结构可分为以下四部分[1-2]:

(1) 文件头。指明了该文件所遵从的 PDF 规范的版本号,它出现在 PDF 文件的第一行。

(2) 文件体。PDF 文件的主要部分,由一系列对象组成。

(3) 交叉引用表。为了能对间接对象进行随机存取而设立的一个间接对象的地址索引。

(4) 文件尾。声明了交叉引用表的地址,即指明了文件体的根对象(Catalog)。

PDF 文件的物理结构如图 1 所示。

图 1 PDF 文件物理结构

图 2 是一个实际的 PDF 文件结构的示意图。

```
%PDF-1.4
...

5 0 obj<</FontDescriptor 6 0 R/W[814 907 500 7716 7716 500]/Type/Font/Subtype/
CIDFontType2/CIDSystemInfo<</Registry(Adobe)/Supplement 2/
Ordering(GB1)>>/WinCharSet 134/BaseFont/#cb#ce#cc#e5/DW 1000>>
endobj

4 0 obj<</DescendantFonts[5 0 R]/Type/Font/Subtype/Type0/Encoding/GBK-EUC-
H/Name/FSec_song/BaseFont/#cb#ce#cc#e5>>
endobj
...
...
...

xref
0 143
0000000000 65535 f
0000373079 00000 n
...

trailer
<</Size 143/ID
[<2067f4c99a995f8f7234412fecd5fe27><634a9040496d42e6377cbda0b5febe08>]/
Info 142 0 R/Root 141 0 R>>
...
```

图 2 PDF 文件结构示意图

PDF 文件的逻辑结构:

一个 PDF 文档是由一些称为“对象”的模块组成的。并且每个对象都有两个数字作为标号,这些对象不需要按照顺序出现在 PDF 文档里面,出现的顺序可以是任意的。

文件尾说明了根对象的对象号,在文件尾中找到关键字/Root,后面的数字就是目录对象(Catalog)的数字编号,示例中的编号为 141 0,在文件中寻找 141 0 的对象。

141 0 obj <</Pages 1 0 R/Type/Catalog>> endobj

该对象表明类型为 Catalog,而且第一页的对象编号为 1 0,对象 1 0 的内容是

1 0 obj <</Type/Pages/Kids[3 0 R 17 0 R 21 0 R 25 0 R 29 0 R 138 0 R]/Count 6>>

该对象表明对象类型是页,而且有 3 0 R 17 0 R 21 0 R 25 0 R 29 0 R 138 0 R 这么多页,中间使用 R 作为页的分割,/Count 表明文件共有 6 页。该文件的第一页是 3 0 对象,3 0 对象中/Contents 表明了该页中内容,同时还表明了字体、页面大小等信息。所以一个 PDF 文件的逻辑结构是一个如图 3 所示的树。

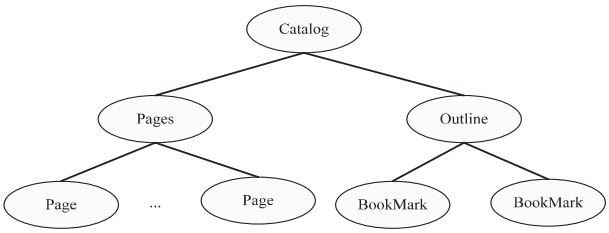

图 3 PDF 逻辑结构图

2 算法思想

2.1 操作 PDF 的方法

对 PDF 文件进行操作的方法有很多,比如通过使用光学字符识别(OCR)软件对 PDF 文件中关键信息进行提取的方法[3],利用 Adobe Acrobat 提供的编程接口编写 Plug-in 插件的方法[4],借助 Acrobat 软件的一个插件 Aerial 或 PDF2TXT 软件来完成的[5]。这些方法对 PDF 文件的操作不是很灵活,对处理后的信息往往还需要手工进行处理,这对于大数据量的操作十分不便。大部分文献采用已有的类库对 PDF 文件进行操作,并且取得了良好的效果[6-8]。

目前,已经存在的对于 PDF 的类库有很多,如 iText、XPDF、PJX、iTextSharp、PDFLib、SharpPDF、PDF-Box 等,但是大多操作类库主要侧重于 PDF 文件的生成操作,对文本进行提取的类库不多,其中 PDFBox 是一个优秀的操作 PDF 文件的 Java 平台类库。

在 PDFBox 中为 PDF 文件中的每个对象建立了相应的类,它们之间的对应关系如表 1 所示。

表 1　PDFBox 中的类与 PDF 对象的对应关系

PDF 类型	PDFBox 类	描述
Array	COSArray	一组有序的数据集合,如 [1,2,3,4,5]
Boolean	COSBoolean	逻辑值 true 和 false
Dictionary	COSDictionary	一组键/值对的集合
Object	COSObject	对象,针对 PDF 文件中的一个完整对象
Stream	COSStream	数据流,可以对 PDF 文件中压缩后的数据流进行操作

通过 COS 模型可以对 PDF 文件进行各种底层的操作。为了操作的简便性,PDFBox 对 COS 模型进行了进一步的封装,提供了 PD 模型,在 PD 模型中提供了文档、页面、字体等类,这样可以更灵活地对 PDF 文件进行操作。

PD 模型与 COS 模型之间的关系如图 4 所示。

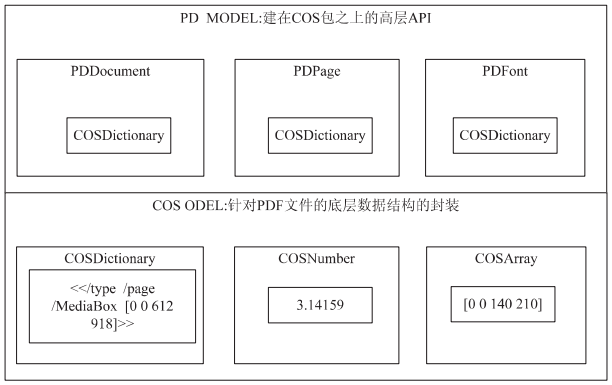

图 4　PDFBox 中 PD 模型与 COS 模型之间的关系

2.2　提取规则

对学术论文信息的提取采用规则的方法[9-14]。

论文标题的提取比较困难,因为各种学术研究发表的论文格式互不相同,标题的位置也千差万别。为了提取到论文的标题,制定以下规则:

(1) 该字符串位于文件的第一页;

(2) 该字符串的字体大小在本页中最大,如果最大字体的对象有多个,选择对象位置的纵坐标最大的对象;

(3) 该字符串的下方不应该有下划线等其他修饰对象。

提取出论文的标题后,开始提取作者,规则如下:

(1) 该字符串的 Position 的 y 值与标题的 Position 的 y 值差最小,且与标题的 y 值差为负值;

(2) 该字符串的字体大小小于标题的字体大小。

如果一个字符串同时满足上述两个条件,则判断其为第一作者。对于非第一作者,参照如下规则:

(1) 字符串的位置的 y 值等于第一作者位置的 y 值;

(2) 该字符串的字体大小和字体类型与第一作者的相同。

对应单位的地址、名称、邮政编码等信息也采用类似的规则进行提取。

对摘要和关键词的提取比较简单,因为几乎所有的学术论文在摘要前都有文字“摘要”或者“Abstract”作为标识,关键词也有类似的标志; 对文章发表日期的提取采用收稿日期作为文章的发表日期,如果文章中没有收稿日期,则在页眉/页脚区域寻找出版日期,并对这两种不同的日期做不同的标记。

对于一篇 PDF 格式的学术论文,进行信息抽取的算法流程如图 5 所示。

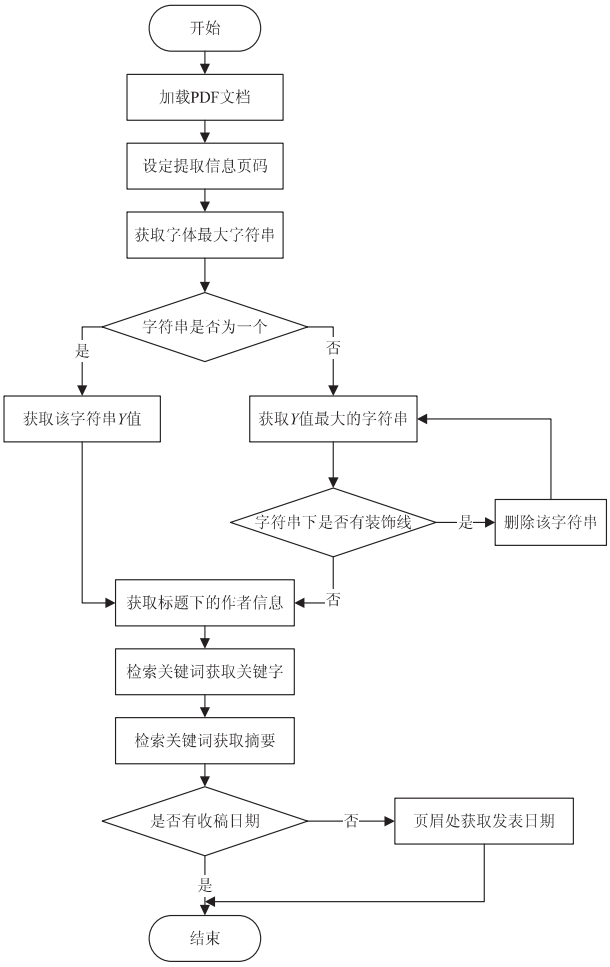

图 5　信息抽取算法流程图

3　测试及结果

通过 CNKI 数据库,搜集了 2 000 篇学术论文,其中中文学术论文 1 250 篇,英文学术论文 750 篇,使用的 PDFBox 版本为 1.8.4,fontbox 为 1.8.4,commons-logging 为 1.13 版。表 2 列出了对论文各部分提取的正确率情况。

表 2　信息提取正确率　　%

	标题	作者	工作单位	摘要	关键词	出版日期
中文论文	88.3	87.4	86.2	100	100	80.2
英文论文	92.7	93.4	92.1	100	100	90.5

(下转第 68 页)

法将无效的语法歧义进行消除，通过分类筛选出候选概念，最后采用 gSpan 算法对频繁子图中的数据进行挖掘，形成 OWL - DL 格式的领域本体描述。在反馈评价阶段，根据前四个环节运算的结果应用了一种自我调整与修正的机制，整个框架能够完整准确地提取概念和关系，有效地自动构建本体，但是在准确率上，随着图上顶点数量的增加有下降的趋势，需要在以后的研究中进一步设计解决。

参考文献:

[1] Ding Shengchun, Jiang Chaonan. Excavating implicit relation based on SWRL [J]. New Technology of Library and Information Service, 2011, V27(3): 68 - 72.

[2] 侯 鑫，张旭堂，金天国，等. 面向知识与信息管理的领域本体自动构建算法 [J]. 计算机集成制造系统，2011，17(1): 159 - 170.

[3] Mao Yuxin, Chen Huajun, Jiang Xiaohong, t al. Domain kn - owledge resource management based on sub - ontology [J]. Computer Integrated Manufacturing Systems, 2008, 14(7): 1434 - 1440.

[4] Gacitua R, Sawyer P, Rayson P, et al. A flexible framework to experiment with ontology learning techniques [J]. Knowledge - Based System, 2008, 21(3): 192 - 199.

[5] Zhao Jianxun, Zhang Zhenming, Tian Xitian, et al. Ontology & its applications in mechanical engineering [J]. Computer Integrated Manufacturing Systems, 2007, 13(4): 727 - 737.

[6] 李志国，冯 永，钟 将，等. 基于 Super - P2P 的分布式知识管理模型 [J]. 计算机科学，2007，34(7): 184 - 186.

[7] 王永贵. 分布式知识管理中的语义交互式框架与方法研究 [D]. 大连: 大连理工大学，2008.

[8] 张海霞，吴 江. 基于语义网的知识管理系统框架设计 [J]. 计算机技术与发展，2006，16(4): 46 - 48.

[9] Dongen S. A cluster algorithm for graphs [R]. New York, NY, USA: ACM, 2000.

[10] 刘 莉，何中市，邢欣来，等. 基于语义角色的中文时间表达式识别 [J]. 计算机应用研究，2011，28(7): 2543 - 2545.

[11] 刘成山，赵捧未. 语义对等网环境下的数字图书馆原型 [J]. 情报杂志，2010，29(6): 110 - 112.

[12] 孙 艳，周学广，付 伟. 基于依存关联分析的情感词扩展 [J]. 北京邮电大学学报，2012，35(5): 90 - 93.

[13] Mihalcea R, Tarau P. TextRank: bringing order into texts [C] //Proceedings of the empirical methods in natural language processing. Berlin, Germany: Springer, 2006: 404 - 411.

[14] Ngomo A C N. SIGNUM: a graph algorithm for terminology extraction [J]. Lecture Notes in Computer Science, 2008, 4919: 85 - 95.

[15] 邬 桐，周雅倩，黄萱菁，等. 自动构建时间基元规则库的中文时间表达式识别 [J]. 中文信息学报，2010，24(4): 3 - 10.

(上接第 63 页)

通过表 2 可以看出，英文论文的提取正确率要高于中文论文的提取情况，主要原因是中文论文中的格式过于多样化，依靠对标题提取规则有时并不能正确地提取到标题，而是提取到页眉的内容，一旦标题提取不正确，作者信息和单位也会提取失败。很多学术论文中并没有收稿日期，尽管程序已经利用页眉或页脚中的日期信息，但是仍然有很多论文甚至连页眉/页脚信息也没有，造成出版日期提取失败。所以提取的规则或者方法还有待进一步完善。

参考文献:

[1] Adobe Systems Inc. PDF reference, Adobe portable document format version 1.4.3nd [EB/OL]. 2001. http://www.adobe.com/suppon/down - loads/product.jsp? product = 44& platform = Windows(Accessed Mar. 8, 2005).

[2] Lovegrove W S, Brailsford D F. Document analysis of PDF files: methods, results and implications [J]. Electronic Publishing Origination Dissemination and Design, 1995, 8(2/3): 207 - 220.

[3] 陈云榕，刘立柱，丁志鸿. PDF 文件中关键信息的提取与组织方法研究 [J]. 计算机工程与设计，2007，28(7): 1688 - 1690.

[4] 李贵林，李建中，杨 艳. 用 Plug - in 实现对 PDF 文件的信息提取 [J]. 计算机应用，2003，23(2): 110 - 112.

[5] 赵 耀. 基于 PDF 文档的数字化学习资源建设 [J]. 临沂师范学院学报，2011，33(6): 125 - 128.

[6] 龙 珑，邓 伟，覃 晓. 绿色网络 PDF 提取系统 [J]. 计算机技术与发展，2014，24(1): 204 - 207.

[7] 张秀秀，马建霞. PDF 科技论文语义元数据的自动抽取研究 [J]. 现代图书情报技术，2009(2): 102 - 106.

[8] 李兰友，陈 立，谢雪莲. 面向 Web 的 PDF 文档构建技术 [J]. 计算机与现代化，2013(12): 184 - 187.

[9] Yuan Fang, Liu Bo, Yu Ge. A study on information extraction from PDF files [C] //Processdings of the 4th international conference on advance in machine learning and cybernetics. Berlin: Spinger - Verlag, 2005: 258 - 267.

[10] 李 强，刘时进. PDF 阅读器的设计与实现 [J]. 计算机工程与设计，2013，31(7): 1635 - 1638.

[11] 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，2002，38(21): 189 - 191.

[12] Chao Hui, Fan Jian. Layout content extraction for PDF documents [C] //Proceedings of document analysis systems. Berlin: Spinger - Verlag, 2004: 213 - 224.

[13] 宋艳娟，张文德. 基于 XML 的 PDF 文档信息抽取系统的研究 [J]. 现代图书情报技术，2005(9): 10 - 13.

[14] 杨道良. 面向对象的中文 PDF 阅读器的设计与实现 [J]. 计算机应用，1999，19(6): 1 - 4.