一种规则与 SVM 结合的论文抽取方法

李雪驹,王智广,鲁 强
(中国石油大学(北京) 地球物理与信息工程学院,北京 102249)

摘 要: 传统 PDF 论文抽取方法主要是单独基于规则的方法或单独基于机器学习的方法,其中基于规则的抽取方法在处理格式固定的数据方面具有明显的优势,通过制定简单的抽取规则即可准确定位并抽取数据;而在处理格式灵活的数据时,则需要制定相当复杂的规则,且不具备对论文格式的适应性,因而明显缺乏机器学习抽取方法的灵活性和准确性。为此,提出了一种基于规则与 SVM 相结合的 PDF 论文抽取方法。该方法充分利用规则方法与机器学习在信息抽取时的优点,在用简单的规则抽取格式固定的信息的基础上,选取样本特征构建训练集,并选择最优的核函数生成 SVM 模型,从而完成基于 SVM 方法的信息抽取。以 SVM 的抽取结果为主体,通过合理利用基于规则抽取的结果并制定适当的规则的方式对该方法进行验证。实验结果表明,该方法在论文元数据和章节标题等信息抽取方面具有较好的效果。

关键词: PDF 论文;规则;支持向量机;样本特征;混合方法;信息抽取
中图分类号: TP301 **文献标识码:** A **文章编号:** 1673-629X(2017)10-0024-06
doi:10.3969/j.issn.1673-629X.2017.10.006

An Extraction Method for Papers via Integration of Rules with SVM

LI Xue-ju,WANG Zhi-guang,LU Qiang
(College of Earth Physics and Information Engineering,China University of Petroleum - Beijing,
Beijing 102249,China)

Abstract: Traditional extraction methods for PDF format papers are mainly based on either rules or machine learning. The extraction method based on rules has obvious advantages in processing fixed format data,which can accurately locate and extract data by making some simple rules of extraction. However it needs fairly complex rules to deal with flexible data and is lack of the adaptability of paper format,which cannot do better than the extraction method of machine learning in terms of flexibility and accuracy. For this,an extraction method for PDF papers via integration of rules with SVM is proposed which makes full use of the advantages of rules and machine learning when extracting information. On the basis of extracting fixed format information via simple rules,the sample characteristics is chosen to build the training set and the optimal kernel function is selected to generate the SVM model for implementation of information extraction based on SVM. By taken extraction results of the SVM as the main body,the verification experiments is conducted based on rules rationally and some appropriate rules made. The experiment results show that it can achieve better results for extracting metadata and chapter headings of PDF papers.

Key words: PDF papers; rules; support vector machine; sample characteristics; hybrid method; information extraction

0 引 言

随着互联网和信息技术的发展,大数据已成为各个领域最热门的名词。面对海量的信息和数据资源,迅速获取其中潜在的、有用的知识是当今数据挖掘的重要方向。学术论文具有强烈的专业性和准确性,论文内的信息和数据在很多专业领域都能发挥极大的作用,能为许多应用技术提供底层的数据支持。因此抽取学术论文中的信息和数据是非常有意义的。

目前国内外的学术论文多以 PDF 格式进行存储,PDF 文档内容抽取主要有两种方式。一种是通过分析 PDF 文档的格式,直接将其中内容抽取出来,进而获取有用的信息和数据,以下简称直接方法[1];另一种是将原 PDF 文档转换成其他文档格式,从而利用抽取中间文档内容的方法抽取 PDF 文档中的内容,再进一步获

收稿日期: 2016-11-27 **修回日期:** 2017-03-14 **网络出版时间:** 2017-07-19
基金项目: 国家自然科学基金资助项目(60803159);国家科技重大专项(2011ZX05005-005-006)
作者简介: 李雪驹(1990-),男,硕士,CCF 会员(200056264G),研究方向为数据挖掘、知识图谱;王智广,教授,博士,CCF 高级会员,通讯作者,研究方向为计算智能、分布与并行计算;鲁 强,副教授,博士,CCF 会员,研究方向为分布式系统、知识工程。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20170719.1113.090.html

取有用的信息和数据,以下简称间接方法[2]。近年来,由于 PDFBox 等开源工具的日益成熟,直接方法得到了广泛应用。

直接方法主要分为基于规则和基于机器学习两大类[3],传统研究多是单独基于规则或机器学习进行 PDF 文档的抽取,以下简称单独方法。尽管在元数据分类抽取等方面取得了较大的成绩,但由于学术论文的格式过于复杂、繁多,上述单独方法在某些情况下的效果并不理想。并且传统研究大多只关注元数据的抽取,没有很好地给出论文的内容结构以及内容中的信息和数据。

由前人的研究可以发现,单独方法在抽取元数据过程中时而效果特别突出,时而效果却很差。为此,提出了一种基于规则与 SVM 相结合的方法。该方法充分发挥了两种方法各自的优点,取得了比单一方法更优的抽取效果,还获得了论文内容、结构等方面的信息数据。

1　PDF 文档的抽取方法

PDF 文档的内容并不是简单的字符串的拼接,它是多个数据对象的组合,因此不能像 WORD 一样抽取文档的内容。目前 PDF 文档内容的抽取主要有直接抽取和间接抽取两类方法。

1.1　直接抽取方法

该方法主要是通过分析 PDF 文档的物理结构和逻辑结构,运用 PDFBox 等开源工具解析 PDF 文档,直接将其中的文本信息和图片抽取出来[4],解析后的 PDF 文档可以通过规则、机器学习以及规则与机器学习相结合等方法进一步抽取有用的信息和数据。

1.1.1　基于规则的抽取方法

基于规则的方法主要采用基于模式识别和模式匹配的模板挖掘技术来实现自由文本的分类抽取。如利用正则表达式从 PDF 文档中抽取首页元数据[5];采用基于层级知识描述框架的 InfoMap 方法抽取引文元数据等[6]。

基于规则的抽取方法易于理解和操作,只要规则制定合理,效果十分明显。但是该方法需要专业人员预先制定一系列规则,而且如果抽取的目标发生变化,则会产生规则不适应的问题。

1.1.2　基于机器学习的抽取方法

机器学习的方法则采用另外一种思路,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,最终达到合理的分类抽取。如采用条件随机场模型抽取多种通用元数据[7];用概率评估模型抽取引文元数据[8];用 SVM 模型抽取论文的元数据[9]等。

机器学习的方法具有较强的适应性,可以处理多种类型的文档,不需要专家提前制定规则,但是这种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以及样本特征的选取。

1.1.3　基于规则和机器学习相结合的抽取方法

规则和机器学习相结合的方法就是在抽取过程中既用到了规则又用到了机器学习。以抽取 PDF 学术论文中的元数据为例,研究发现,基于规则的抽取方法在处理某些元数据时的效果要优于机器学习方法,比如参考文献、摘要及关键词的抽取;然而在抽取文章标题、作者信息等元数据时的效果却不如基于机器学习的方法。这主要是因为参考文献等元数据通常会满足一定的格式,并且基本不会改变,而文章标题等元数据则不具备这样的规则性。与此同时,有些关键信息需要极其复杂的规则才能获取,而用机器学习的方法则可以较轻松地得到。

基于前面的分析,分别用规则和机器学习抽取各自适合的信息和数据,再将它们统一起来,能够显著地提高抽取结果;并且对于机器学习不准确的地方,也可以通过适当的规则进行修正以提高抽取的准确率。这种方法具有较强的适应性,同时能够减少规则设计的复杂性,只需要制定一些简单规则,基本可以解决 PDF 文档抽取过程中的各类问题。

1.2　间接抽取方法

这种方法主要是将原 PDF 文档转换成其他文档格式,从而利用抽取中间文档内容的方法抽取 PDF 文档中的信息。已有方法包括基于 XML 的 PDF 文档信息抽取、基于 XSLT 的 PDF 论文元数据的抽取[10-11]。随着 OCR 技术的提高,将 PDF 文档的内容转换成 OCR 扫描的图片进行信息抽取也得到了越来越多的重视。

2　混合方法抽取 PDF 学术论文

PDF 学术论文的元数据主要包括文章标题、作者信息、摘要、关键词以及参考文献等。不难发现,摘要、关键词以及参考文献的出现都会有一个明显的标志,例如“摘要”、“Abstract”、“关键词”等。因此采用基于规则的方法可以简单、迅速地定位并抽取这些内容。对于文章标题及作者信息等元数据,由于它们的出现相对灵活,没有明显的标志,所以机器学习的方法能够更准确地抽取这部分元数据。再来研究文章的内容信息,众所周知除了上述论文的元数据,文章内容同样包含了许多重要的信息和数据。例如论文各章节的标题及子标题,论文表格内的信息和数据等。提出的方法不但准确地抽取了 PDF 论文基本的元数据,而且还抽取了论文的章节标题等重要的内容信息。

对提出的混合方法的核心思想、方法流程进行介

绍,如图1所示。其中曲边四边形表示文档、文件,矩形表示必须处理的过程,平行四边形表示数据,椭圆形表示注释。

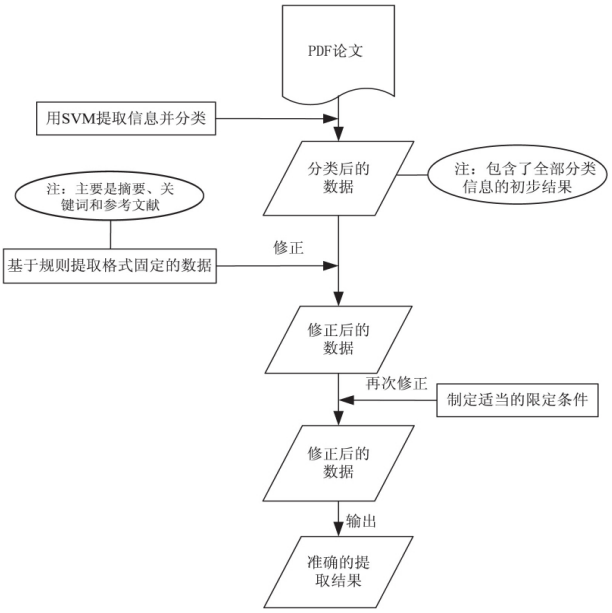

图1 混合方法的具体流程

具体过程如下:先利用生成的SVM训练模型对PDF论文进行分类,初步得到一个分类结果,包括文章标题、作者信息、正文内容、章节标题、页脚页眉以及摘要、关键词和参考文献;接着利用基于规则抽取到的摘要、关键词以及参考文献去修正SVM得到的结果;然后再按照论文格式等限定条件去适当修正其他不合理的分类信息,最终得到相对准确的各类信息和数据。

2.1 基于规则抽取方法的实现

基于规则的抽取方法主要用来抽取PDF论文内格式固定的信息和数据,一般指摘要、关键词和参考文献。PDFBox是一个很好的开源PDF文档解析类库,首先用PDFBox解析PDF论文,然后利用其接口将PDF的内容流按照论文显示的行去存储。每一行都包含了这一行的位置信息、字体信息等重要内容。接下来制定规则分别去抽取论文的摘要、关键词以及参考文献。

这三类元数据的抽取方法大致相同,都是基于字符串匹配的方式。具体方法如下,按行遍历所有的论文内容,分别寻找行首带有“摘要”(或ABSTRACT、Abstract等)、“关键词”(或关键字、主题词、Keywords等)、“参考文献”(或REFERENCE、Reference等)的行,确定这些行的位置。值得一提的是,摘要和关键词多出现在论文的首页,参考文献多出现在文章的结尾。如果能够找到上述三行的位置,即说明此论文包含摘要、关键词和参考文献的内容。

此时摘要所在行与关键词所在行之间的内容是论文的摘要部分,且摘要各行之间的字体大小应该是相同的(在误差允许范围内);关键词所在行的内容是论文的关键词部分,由于关键词可能不止一行,所以还应该再向下判断一至两行(关键词一般只有1~3行),判断方法与摘要相同,用关键词各行之间的字体大小来判断该行是否为关键词,最后得到正确的关键词内容。参考文献部分的抽取,从出现该字符串的下一行开始,逐行比较各行的字体大小,连续的字体大小相同的行就组成了论文的参考文献部分。

如果无法全部找到前文所说的“摘要”、“关键词”以及“参考文献”的行,那么说明该文章缺少其中某些部分的内容,即是说缺少哪一行就不存在哪一种元数据。此时要充分利用同一种元数据相邻行的字体大小相同、行间距无明显变化等方法进行划分,抽取对应种类的元数据。

2.2 基于SVM抽取方法的实现

用规则抽取PDF论文的元数据主要是通过数据的位置和字体大小来判断分类,然而很多时候无法轻易地对数据进行分类。例如有时解析后的PDF文档,内容流中的字体大小都是0,这时就无法从这一特征量判断字体的大小。因此还需要考虑字符的宽度、高度、横纵坐标的比例等特征量,综合起来判断实际显示在文档中的字体大小。这里需要考虑的特征量越多,制定的规则就越复杂,并且可能存在的误差也越大。这时应该采用机器学习的方法抽取数据。

PDF论文的信息抽取实际上也是一种分类问题,由于SVM在解决分类和回归问题方面性能显著,具有良好的理论证明,并且可以很好地支持小样本,因此选用SVM作为机器学习的核心方法。

2.2.1 SVM的特征选取

用SVM抽取PDF学术论文本质上就是将PDF论文分类,这是一个多分类问题。大体上可以将PDF学术论文分为以下几类,分别是文章标题、作者信息、摘要、关键词、正文内容、各章节标题、参考文献以及页脚页眉等。针对上面这些信息和数据在PDF文档中的特性,合理地选取训练样本的特征。分析论文结构不难发现,区分上面分类的主要因素就是位置和字体,因此要在内容流中寻找与位置和字体相关的样本特征。

利用SVM模型,将论文中的每一行进行分类。由于PDF论文的每一行都包含了反映其位置和字体的特征,行可以很好地表现PDF论文的内容和结构,并且与区域(块)相比,行更能细化这些特征,增强分类的准确性,区域(块)也是由多个行组成的;与此同时,还能更好地与基于规则的方法相结合。因此,采用以行为基本单位,运用SVM模型进行分类的方法。

训练SVM模型,最重要的是把论文行转换成SVM的特征向量。经过解析后的PDF内容流按行存储,每

行都包含了位置和字体等信息,针对这些信息,合理选择特征向量。

选择行的位置特征。一般来说,同一行的每个字符的纵坐标是相同的,选择每一行的第一个字符的横坐标 XDirAdj、纵坐标 YDirAdj 及最后一个字符的横坐标 XDirAdj 作为特征向量。首尾两个横坐标表示这一行的长度,加上纵坐标基本上就能够确定该行在 PDF 文档中的位置。

选择行的字体特征。多数情况下,同一行的字体特征是相同的,因此选择第一个字符的字体大小 FontSize 来代表这一行的字体大小。然而实验发现,有时 FontSize 在整篇文章中的值都是 0,单靠 FontSize 一个特征向量并不能反映字体的大小,还要考虑字体高度 HeightDir、字体宽度 WidthDirAdj、字体横坐标比例 XScale、字体纵坐标比例 YScale 以及字体 Pt 尺寸 FontSizeInPt。将上述参数作为表示这一行字体大小的特征向量可以很好地反映这一行的分类特征。

因为有些情况下还需要考虑行的字符个数以及该行所处的 PDF 文档的页码,比如文章标题、作者、章节标题、正文内容等在每一行的字数都会有一定差别,并且文章标题、作者、摘要、关键词等多出现在 PDF 论文的首页,所以每一行的字符个数和所处的页码也可以作为样本特征。

此外,论文行还包含了前后行间的距离、字体格式、字体方向、字体间距等特征。将上述特征分成几组训练 SVM 模型,测试结果见表 1。

表 1　不同特征向量的 SVM 模型的简单对比

类别	准确率/%
A	90.6
B	83.7
C	87.7
D	90

表 1 中类别 A 选择了每一行第一个字符的横坐标 XDirAdj、纵坐标 YDirAdj、字体大小 FontSize、字体高度 HeightDir、字体宽度 WidthDirAdj、字体横坐标比例 XScale、字体纵坐标比例 YScale、字体 Pt 尺寸 FontSizeInPt、最后一个字符的横坐标 XDirAdj、该行的字符个数以及所处的 PDF 文档的页码共 11 个特征向量; 类别 B 选择了每一行第一个字符的横坐标 XDirAdj、纵坐标 YDirAdj、字体大小 FontSize、最后一个字符的横坐标 XDirAdj、该行的字符个数以及所处的 PDF 文档的页码共 6 个特征向量; 类别 C 选择了每一行第一个字符的横坐标 XDirAdj、纵坐标 YDirAdj、字体大小 FontSize、字体高度 HeightDir、字体宽度 WidthDirAdj、字体横坐标比例 XScale、字体纵坐标比例 YScale、字体 Pt 尺寸 FontSizeInPt、最后一个字符的横坐标 XDirAdj 共 9 个特征向量; 类别 D 选择了每一行第一个字符的横坐标 XDirAdj、纵坐标 YDirAdj、字体大小 FontSize、字体高度 HeightDir、字体宽度 WidthDirAdj、字体横坐标比例 XScale、字体纵坐标比例 YScale、字体 Pt 尺寸 FontSizeInPt、最后一个字符的横坐标 XDirAdj、该行的字符个数以及所处的 PDF 文档的页码、前后行间的距离、字体方向、字体间距共 14 个特征向量。

实验随机选用了相同的标注好的 1 000 个样本行训练模型,并随机选用另外的 350 个样本行进行测试,未经过参数调优,选用相同参数的 RBF 核后粗略地得到表 1 所示的结果。

由表 1 可知,类别 A 的准确率相对高些,因此最终选取了每一行第一个字符的横坐标 XDirAdj、纵坐标 YDirAdj、字体大小 FontSize、字体高度 HeightDir、字体宽度 WidthDirAdj、字体横坐标比例 XScale、字体纵坐标比例 YScale、字体 Pt 尺寸 FontSizeInPt、最后一个字符的横坐标 XDirAdj、该行的字符个数以及所处的 PDF 文档的页码这 11 个特征向量作为 SVM 模型的样本特征。

根据 PDFBox 解析后的内容流,对照 PDF 学术论文人工标注训练集和测试集,训练样本的分类包括文章标题、作者信息、正文内容、章节标题、页脚页眉,以及摘要、关键词和参考文献。

2.2.2　SVM 核函数的选取

完成训练样本后要选择合适的核函数来训练模型,选用 LIBSVM 生成训练模型。LIBSVM 是台湾大学林智仁教授开发的一套开源的 SVM 软件包,它提供了丰富的工具以及多种语言的源码。

由于训练集的样本特征远远少于样本数量,应该选择非线性核函数[12]。常用的非线性核函数主要有多项式核、RBF 核、SIGMOD 核以及混合核[13]。利用 LIBSVM 软件包内提供的工具和源代码,用网格搜索、交叉验证等方法分别找到满足上述核函数的最优参数 C、g、d 和 coef0 以及混合核的权值。需要说明的是,有些核函数并不需要上面全部的参数,根据不同的核函数找到不同的最优参数。然后利用训练集和测试集训练 SVM 模型,对比分析不同核函数的性能,最终选取最优的核函数及其训练模型。

2.3　混合方法的具体实现

利用前面训练好的 SVM 模型对每一篇 PDF 论文的内容进行分类抽取,得到初步抽取结果,如图 2、图 3 所示。

这相当于将整篇文章转换成对应的 SVM 模型的抽取特征,然后进行分类。此时的抽取结果包含了该篇论文的全部分类信息,例如文章标题、作者信息、摘

要信息、关键词信息、文章内容信息、参考文献以及页脚页眉等。图2每行都有12列，第1列表示这一行的分类结果。在这一列“0”表示文章标题，“1”表示作者信息，“2”表示文章摘要，“3”表示关键词及分类号，“4”表示正文内容，“5”表示页眉页脚，“6”表示正文的章节标题，“9”表示文章的参考文献等；第2～12列则表示SVM模型的11个样本特征，这里对每一列的样本特征，都按照规范进行了归一化处理。图3显示了论文内容的按行抽取，每行都能对应图2所示的特征向量。每行最后的三个数字分别代表这一行内容的类别(即分类结果)，所处的PDF文档的页码以及在该页的行数。例如“曲江秀，高长海，查明 = = = 1 0 4”这一行，“1”表示这一行的内容是作者信息，“0”表示这一行位于PDF文档的第一页，“4”表示这行是这一页的第五行，其余内容依此类推。

```
5 1:0.05479452 2:0.11334997 3:0.40789473 4:0.40789476 5:0.4078949
6:0.40789473 7:0.40789473 8:0.44444445 9:0.3449651 10:0.25438598 11:0.0
0 1:0.08675799 2:0.18667172 3:1.0 4:1.0 5:1.0 6:1.0 7:1.0 8:1.0 9:0.8673978
10:0.16666667 11:0.0
1 1:0.4155251 2:0.22604822 3:0.65789473 4:0.65789473 5:0.6578948 6:0.65789473
7:0.65789473 8:0.6666667 9:0.5802592 10:0.078947365 11:0.0
1 1:0.34018266 2:0.24369974 3:0.35526314 4:0.35526317 5:0.35526296
6:0.35526314 7:0.35526314 8:0.3888889 9:0.6440678 10:0.23684211 11:0.0
2 1:0.047945205 2:0.27900282 3:0.4210526 4:0.42105263 5:0.4210527 6:0.4210526
7:0.4210526 8:0.44444445 9:0.90129614 10:0.42982456 11:0.0
2 1:0.05022831 2:0.29801214 3:0.40789473 4:0.40789476 5:0.4078949
6:0.40789473 7:0.40789473 8:0.44444445 9:0.89930207 10:0.4385965 11:0.0
```

图2 用SVM模型得到的抽取特征及分类结果

```
文章编号：1000—2634(2008)03—0024—05  ===5 0 2
柴北缘冷湖一南八仙构造带油气运移通道研究  ===0 0 3
曲江秀，高长海，查明  ===1 0 4
(中国石油大学地球资源与信息学院，山东东营257061)  ===1 0 5
摘要：柴达木盆地北缘冷湖一南八仙构造带的油气勘探实践表明，研究区
内主要有三种油气运移通道：连通的砂体  ===2 0 6
和不整合面通常是油气侧向运移的通道，断层是油气进行垂向运移的重要
通道。综合地质、地震、测井和分析化验等  ===2 0 7
资料，应用现代油气成藏理论、地球物理学等理论和方法，对研究区的油
气运移通道及其与油气成藏的关系进行了研  ===2 0 8
究。研究结果表明，油气运移通道类型及其空间组合方式是控制油气运
移、聚集以及成藏的主要因素。探讨了该区  ===2 0 9
油气运移输导层(不整合面、连通砂体、断层及三者的组合形式)对油气运
移、聚集、成藏的控制作用及其与油气藏形  ===2 0 10
成和分布的关系。为勘探实践提供指导。  ===2 0 11
关键词：不整合面；连通砂体；断层；油气成藏；柴达木盆地  ===4 0 12
```

图3 用SVM模型得到的论文内容的分类结果

图2和图3反映了PDF论文经过SVM模型分类后的初步抽取结果。通过观察可以发现，这个抽取结果还存在一定的分类错误。例如图3，行尾数字为12，行首为“关键词”那一行，这一行SVM分类得到了错误的分类结果，将“关键词”误识别成了正文，因此这一行正确的分类结果应该为“3”而不是“4”。

由前文论述可知，基于规则的抽取方法在抽取论文的摘要、关键词和参考文献等数据时具有明显的优势，所以利用基于规则抽取的格式固定的数据去替换SVM模型的抽取结果。

用设计好的规则按行抽取论文的摘要、关键词和参考文献，分别记录好它们所处的位置，主要是每一行所处的页码和在该页的行数等。为了方便，后文用(页码，行数)表示论文每一行的内容；然后利用这些页码和行数，去修正SVM分类的结果，即在SVM的分类结果中，找到相应的页码和行数，然后将这一行的类别强制替换成基于规则抽取到的结果。例如在图3中，SVM模型的分类结果将(0，12)行的内容识别成了“正文内容”，而基于规则的方法则将(0，12)行的内容识别为“关键词”，将SVM分类结果中的(0，12)行的类别“正文内容4”修改为“关键词3”。对于摘要，关键词和参考文献都按照上述方法进行处理，得到修正后的分类结果。如果利用规则无法得到“摘要”或“关键词”或“参考文献”的数据，则无需修改SVM模型的分类结果。

对于修正后的分类结果还要制定一些限定条件进行二次修正，以确保最终输出的分类结果的准确性。具体的限定条件如下：(由于多数中文论文都包含中文和英文的标题、作者信息、摘要和关键词，这里只抽取其中文的标题、作者信息、摘要和关键词；若是英文论文则无此说明。)

(1) 文章标题“0”只能位于PDF文档的首页，并且在首页的上半部分，最多只能有两组字符串(中文标题和英文标题)，其他页面均不能再出现“0”的分类结果；

(2) 作者信息“1”位于PDF文档的首页，多在文章标题后面出现，其他页面均不能再出现“1”的分类结果；作者信息内包含了各个作者的姓名，所属单位以及部分简介，需要制定简单的规则分别获取上述信息。一般来说，每个作者的中文姓名不会超过4个字，并且所属单位都会用“(”扩起来，分别得到作者姓名和所属单位后，一般剩下的内容为作者简介；

(3) 参考文献“9”位于PDF文档的最后部分，一般在文档的最后一页或最后两页，其他页面均不能出现“9”的分类结果；

(4) 章节标题“6”也要加入一些限定条件，章节标题要在关键词后面出现，属于正文部分，字数一般不超过15，并且在抽取到的字符串中不存在逗号、引号、句号等符号，有时在字符串首部可以出现“数字”或“数字+点号”或“数字+顿号”的组合，例如“1”、“一”、“1.”、“一、”等；

(5) 将不满足上述限定条件的分类结果的类别强制修改为正文内容“4”。

上述限定条件基本上是通用的，能够满足绝大部分的论文格式和内容，但不是绝对的。可以根据不同的情况、不同的需求适当修改。

完成上述多个步骤后，最终会得到相对准确的PDF论文分类抽取结果，至此便完成了混合方法的实现。

3　测试结果与分析

表 2 给出了选定 C 和 g 后不同的核函数的分类结果。

表 2　SVM 不同核函数的分类结果

核函数	C	g	d	coef0	Accuracy/%
线性核	100				79.32
RBF 核	100	0.75			92.41
多项式核	100	0.75	3	3	91.98
SIGMOD 核	100	0.75	3	3	40.93

由表 2 可以看出,使用线性核测试集的准确率只有 79.32%,远小于 RBF 核与多项式核的结果,进一步证明了文献[10]总结的结论,理应选用非线性核函数。又因为 SIGMOD 核的测试效果很不理想,所以主要考虑 RBF 核与多项式核。

深入对比分析 RBF 核与多项式核,这两种核函数都能取得良好的测试结果,但是随着参数的优化,多项式核的训练时间大大超过了 RBF 核的训练时间,而测试集的结果相差不大,因此选择参数调优后的 RBF 核作为该混合方法中 SVM 的核函数。

随机测试了 348 篇 PDF 学术论文,得到的对比结果如表 3 所示。

表 3　三种方法抽取信息的准确率

方法名称	文章标题	作者信息	摘要	关键词	参考文献	章节标题	平均值
规则方法	0.807 5	0.773	0.933 9	0.925 2	0.942 5	0.75	0.855 4
SVM 方法	0.916 7	0.824 7	0.856 3	0.884 8	0.836 2	0.813 2	0.855 3
混合方法	0.922 4	0.839 1	0.933 9	0.925 2	0.910 9	0.839 1	0.895 1

注: 规则方法表示单独基于规则的抽取方法,该方法按照文献[14]介绍的算法思想设计实现; SVM 方法表示单独基于 SVM 的抽取方法; 混合方法则表示基于规则和 SVM 相结合的抽取方法。

表中分别列出了文章标题、作者信息等六种重要数据信息的抽取结果,从结果上看基于规则的方法在抽取摘要、关键词及参考文献方面表现突出,而基于 SVM 的方法在抽取文章标题、作者信息和章节标题方面表现突出。混合方法同时涵盖了两种方法的优势,基本上在各类数据的抽取结果都是最优的,然而抽取参考文献的结果却略逊于规则方法,这主要是由于部分论文格式混乱,在一篇文章中会穿插两篇文章的信息,使得用规则去修正 SVM 分类极为困难,与此同时 SVM 分类也会产生一部分规则难以修正的结果,因此这部分的抽取结果稍差。

除了上述六种信息,混合方法还准确地抽取了论文的页脚页眉、正文内容等关键信息,准确率都在 85% 以上。从整体上看,混合方法取得了较好的抽取效果。

4　结束语

传统方法在抽取论文信息时还存在一定不足,为了更好地抽取 PDF 论文内的关键信息,提出了一种基于规则和 SVM 相结合的 PDF 论文抽取方法。该方法以 SVM 为主体,合理利用规则去修正,最终得到了更准确的抽取结果。与传统单独基于规则或机器学习的方法相比,明显提高了抽取效果,而且还准确地得到了章节标题、页眉页脚等关键信息。

由于 SVM 的训练样本无法包含全部格式的 PDF 论文,所以生成的模型会存在一定的局限性,针对某些特殊格式的 PDF 论文效果会很差; 同时测试论文的数量偏少,也会影响实验结果。在进一步优化训练模型、增加测试论文数量后,要继续深入研究正文内关键信息和数据的抽取,因此准确抽取图片与表格内的数据将是接下来研究的重点。

参考文献:

[1] 李　珍,田学东. PDF 文件信息的抽取与分析[J]. 计算机应用,2003,23(12):145-147.

[2] 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.

[3] 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.

[4] 王晓娟,谭艳龙,刘燕兵,等. 基于自动机理论的 PDF 文本内容抽取[J]. 计算机应用,2012,32(9):2491-2495.

[5] 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.

[6] Day M Y,Tsai R T H,Sung C L,et al. Reference metadata extraction using a hierarchical knowledge representation framework[J]. Decision Support Systems,2007,43(1):152-167.

[7] Yu J,Fan X. Metadata extraction from Chinese research papers based on conditional random fields[C]//Fourth international conference on fuzzy systems and knowledge discovery. [s. l.]: IEEE,2007: 497-501.

[8] Giles C L,Bollacker K D,Lawrence S. CiteSeer: an automatic citation indexing system[C]//Proceedings of the third ACM conference on digital libraries. [s. l.]: ACM,1998: 89-98.

[9] 欧阳辉,禄乐滨. 基于 SVM 的论文元数据抽取方法研究[J]. 电子设计工程,2010,18(5):4-7.

[10] 宋艳娟,李金铭,陈振标. 基于 XSLT 的 PDF 信息抽取技术的研究[J]. 计算机与数字工程,2008,36(5):156-159.

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

[12] Chang C C,Lin C J. LIVSBM: a library for support vector machines[EB/OL]. 2013. http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf.

[13] 赵丽琴. 混合核支持向量机在地铁客流预测中的应用研究[D]. 兰州: 兰州交通大学,2015.

[14] 牛永洁,薛苏琴. 基于 PDFBox 抽取学术论文信息的实现[J]. 计算机技术与发展,2014,24(12):61-63.