文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

基于自动化解析技术的专利申请管理系统设计与实现

付晓明
国能新疆化工有限公司
DOI:10.12238/acair.v3i2.13559

[摘　要] 随着全球专利申请量的激增,多国专利局(如中国EAC系统、国际CPC分类系统)导出文件的异构性成为专利申请管理系统(PMS)高效运行的瓶颈。传统人工解析方式效率低、错误率高,亟需自动化解决方案。本文主要论述如何开发一种能够智能解析EAC与CPC导出文件的专利申请管理系统,解决多格式兼容性问题,提升数据转换效率与准确性,支持跨国专利申请流程。本文以JAVA编程语言为例实现自动解析EAC与CPC发文通知书,一键导入专利申请管理系统自动创建及更新专利信息、缴费信息、发文通知书等关键信息。
[关键词] 异构数据兼容；XML解析技术；多格式动态适配；专利申请流程管理自动化
中图分类号：O175.11　文献标识码：A

Design and Implementation of Patent Application Management System Based on Automated Analysis Technology

Xiaoming Fu
CHNENERGY XINJIANG CHEMICAL CO.,LTD.

[Abstract] With the surge in global patent applications, the heterogeneity of exported files from multiple patent offices (such as the Chinese EAC system and the international CPC classification system) has become a bottleneck for the efficient operation of patent application management systems (PMS). Traditional manual parsing methods have low efficiency and high error rates, and there is an urgent need for automated solutions. This article mainly discusses how to develop a patent application management system that can intelligently parse EAC and CPC export files, solve multi format compatibility issues, improve data conversion efficiency and accuracy, and support cross-border patent application processes. This article takes JAVA programming language as an example to achieve automatic parsing of EAC and CPC publication notices, and one click import of patent application management system to automatically create and update key information such as patent information, payment information, and publication notices.
[Key words] Heterogeneous data compatibility; XML parsing technology; multi format dynamic adaptation; patent application process management; automation

引言

在专利审查与管理过程中,EAC(中国专利电子申请系统)和CPC(国际专利分类系统)的发文通知书是核心文件,承载着审查意见、法律状态、费用缴纳等关键信息。然而,由于两者在文件结构、数据格式、编码规范上的显著差异,传统人工解析方式效率低下,且易因嵌套层级复杂、非结构化文本等因素导致信息提取错误。为解决上述问题,本文提出一种基于多模态数据解析引擎的智能处理方案,通过分层架构设计与AI融合技术,实现EAC/CPC文件的自动分类、结构化提取与逻辑校验。

1 EAC/CPC发文通知书文件结构特征分析

首先,实现一键导入的前提是需要分析EAC和CPC发文通知书文件特征。从中国专利电子申请系统(EAC)导出的文件通常基于XML或固定格式文本,包含结构化标签(如<通知书名称>,<申请号>,<发文日期>等)和嵌套层级,每个通知书的标签信息都不一样。国际专利分类系统(CPC)导出的文件多采用XML或JSON格式,遵循WIPO标准,包含分类号(如G06F40/205)、主权项关联关系和多语言支持字段。其中关键字段包括主分类号、引用文献DOI、权利要求树形结构、多语言摘要,详细对比这两种发文的格式标准、嵌套深度、非结构化内容及编码规范见表1 EAC与CPC导入文件结构特征对比。

分析发现EAC和CPC导出的文件结构规律,通过标签(XML)或键值对(JSON)定义数据边界,部分字段需跨节点关联解析(如权

利要求引用关系)。专利申请完成后收到的通知书包括办理登记手续通知书、受理费减通知书、受理缴纳申请费通知书三种类型,每种类型的格式都不尽相同,嵌套深度和结构化标签详细信息见中国专利电子申请系统EAC提供的schema文件,下图以办理登记通知书为例解析说明书的含义：

表1　EAC与CPC导入文件结构特征对比

特征	EAC	CPC
格式标准	中国国知局自定义 XML Schema	WIPO 国际标准 XML Schema
嵌套深度	浅层(2-3 级)	深层(4-5 级,含权利要求树)
非结构化内容	审查意见正文(自由文本)	主权项逻辑表达式(半结构化)
编码规范	GB18030/UTF-8	UTF-8/Unicode

表2 办理登记手续通知书标签含义

通知书名称	节点名称					层级	中文
办理登记手续通知书	cn_notice_info					0	通知书信息根节点
		notice_name				1	通知书名称
		notice_sent				1	
		notice_sent	notice_sent_date			2	发文日期
		application_number				1	申请号
		pay_deadline_date				1	缴费截止日期
		fee_info_all				1	所有费用信息
		fee_info_all	fee_info			2	费用明细
			fee_info	fee		3	
			fee_total		fee_name	4	费用种类
					fee_amount	4	金额
							费用金额总计(在受理缴纳申请费通知书、受理费减通知书中)
			fee_paid			2	已缴费用
			fee_payable			2	应缴费用
			annual_year			2	缴纳年费年度
			cost_slow_flag			2	减缓标记

2 设计与实现

如何使用AI技术对文件类型进行判断和解析,本系统设计了一个多模态数据解析引擎。这个引擎是处理异构数据源(如文本、图像、结构化文档)的核心技术组件,其核心目标是将不同模态、不同格式的输入数据转化为统一的结构化表示。

多模态解析引擎采用分层处理架构,通过模块化设计实现高扩展性与容错性。一共包含文件导入、格式检测与适配、模态解析、跨模态特征融合、结构化存储等5层处理架构,具体功能如下：

2.1文件导入层

人机交互界面,支持单个、批量文件导入,单个文件可直接选择上传文件类型(发文通知书、证书文件)、申请人(非必填)、申请号(必填)信息确认后调用格式检测与适配层处理,而批量上传功能无需加载任何信息,只将多个文件上传到服务器上后直接调用格式检测与适配层模块进行处理。

2.2格式检测与适配层

格式进行检测包含两个步骤,首先须符合压缩文件格式,其次是基于压缩文件名称结构特征判断文件类型。调用专用解析层接口解压后,解压过程中因GA开头的压缩文件嵌套层级较深,需要使用递归算法逐层解压,查找关键文件并进行解析。通过文件头标识自动识别格式,例如GA开头文件(通知书ID)和非GA开头纯数字(ZIPBID)的证书文件或申请回执。

```
//读取压缩包的文件,并解压到服务器对应目录上
public void readZip(UploadFile upfile, UserSession usi, File file, String uploadPath, String packName, String importZipUrl,String importFileName)throws Exception{
    BufferedOutputStream bos=null;
    FileInputStream fis=new FileInputStream(file);
    BufferedInputStream bis=new BufferedInputStream(fis);
    ZipInputStream zis=new ZipInputStream(bis);
    ZipEntry entry;
    File Fout=null;
    String filePath= uploadPath+"/undozip/"+packName;//解压后路径
    while((entry=zis.getNextEntry())!=null){//递归读取文件条目,只要不为空,就进行处理
    int count ;
    byte date[]=new byte[BUFFER];
    if(entry.isDirectory()){
    continue;//如果条目是文件目录,则继续执行
    }else{
    Fout=new File(filePath,entry.getName());
    if(!Fout.exists()){
    (new File(Fout.getParent())).mkdirs();}
    bos=new BufferedOutputStream(new FileOutputStream(Fout));
    while((count=zis.read(date))!=-1){
    bos.write(date,0,count);}
    bos.flush();
    bos.close();
    }
    }
    zis.close();}
```

2.3模态专用解析层

模态专用解析层的接口为多模态,可处理结构化数据、非结构化数据、图像等文件。EAC的schema包括专利申请主文件(EAC_Application.xsd)、审查意见通知书(EAC_ReviewOpinion.xsd)、费用缴纳通知书(EAC_Fee.xsd)等类型,schema明确了数据结构的定义,如下几种数据类型根据这个定义可以进行动态匹配。

(1)结构化数据(XML)：建立预定义规则库,使用XPath/JSONPath提取字段,并根据文件类型查找对应的schema进行校验,匹配规则库的KEY来执行对应的处理方法。因部分结构化数据具备嵌套属性,需要使用递归解析算法提取相关内容并建立关联关系。

(2)非结构化文件或图像：如审查意见、权利要求、证书文件等,由OCR(Tesseract)完成识别后,采用规则引擎+机器学习,通过预定义正则模板提取段落,后用机器学习训练模型识别术语,本系统最初通过读取list.xml文件关键信息与专利信息进

行关联,通知书直接显示图片而不解析成文字,后改进使用Tesseract完成OCR识别,将文字提取出来后保存。

```
public void readUndoZipNotice(UploadFile upfile, UserSessionInfo usi,String filePath,String importZipUrl,
String importFileName,List<TFileDetail>tfdlist) throws Exception {
int fileType=upfile.getApplyType();//上传文件类型
File pack=new File(filePath);
File[]packs=pack.listFiles();
TFile tfileR=null;
for(int i=0;i<packs.length;i++){
File file=packs[i];
if(file.getName().indexOf("list.xml")!=-1){
//文件list.xml,主要提取专利基本信息。
//分析list.xml文件
tfileR = analysisXml(upfile, usi, file, importZipUrl, importFileName);
if(tfdlist.size()>0){//需要将文件明细与文件与案件信息关联起来
for(TFileDetail tfd:tfdlist){
tfd.setFileId(tfileR.getId());
fileDetailDao.save(tfd);
}
tfdlist.clear();//每个list.xml对应一个发文通知书,处理完本发文通知书后要清空图片文件list
tfileR=null;
}
}else if(file.isDirectory()&& (file.getName().indexOf("GA")!=-1)){
String filePathSon=file.getPath();
readUndoZipNotice(upfile,usi,filePathSon,importZipUrl,importFileName,tfdlist);
}elseif(file.getName().indexOf(".tif")!=-1||file.getName().indexOf(".jpg")!=-1||
file.getName().indexOf(".png")!=-1||file.getName().indexOf(".gif")!=-1||
file.getName().indexOf(".bmp") !=-1){
savePictrue(upfile, tfdlist,tfileR, file);}
}
}
public void OCRConvert{//图像识别方法
Tesseract tesseract=new Tesseract();//创建Tesseract实例
try {
tesseract.setDatapath("path/to/unzipdata");// 设置数据路径
tesseract.setLanguage("eng");//设置语言
File imageFile=new File("path/to/your/image.png");//解压后通知书图像文件
String result=tesseract.doOCR(imageFile);//进行OCR识别
System.out.println("识别结果:"+result);//输出识别结果
}catch (TesseractException e){
System.err.println("OCR识别出错:"+e.getMessage());
}
}
```

2.4跨模态特征融合层

模态解析工作完成后,就需要建立不同模态数据的关联映射(如PDF中的“权利要求1”文本与XML中的<claim>节点对齐等),在融合的过程中,同时进行逻辑校验和数据修复,例如有的通知书中缺失必要字段,需要进行上下文推测,从申请号反推申请日；驳回决定必须关联法律依据等,保证抓取的内容的逻辑完整性。

通过上述设计,可实现跨EAC/CPC系统的发文通知书高效解析,满足专利申请管理系统对多源异构数据兼容性的核心需求,实验表明,该系统对EAC文件的解析效率较传统人工操作提升15倍,某代理机构使用系统后,审查意见处理时间从3小时/件缩短至12分钟/件。

[参考文献]

[1]王磊.基于多模态融合的欧亚经济联盟(EAC)认证文件智能解析系统设计[J].计算机应用研究,2022(8),2345-2352.

作者简介：

付晓明(1978--),女,汉族,辽宁省人,本科,工程师,从事的研究方向：计算机及应用。