文章编号：2096-1472(2018)-01-15-03
DOI:10.19644/j.cnki.issn2096-1472.2018.01.004

文档生成技术研究与应用

邵欣欣，张明会，高梓峻
(大连东软信息学院软件工程系，辽宁 大连 116023)

摘　要：本文的在线生成在当前的信息系统中有广泛的应用，基于现有的POI和Itext等API无法实现的问题进行扩展。文中研究了文档的直接生成中的难点问题，包括复杂表格的生成、单元格的合并等；还研究了对Word和Excel的内容进行替换的文档生成方式，总结了适用于Word和Excel文档的最优生成方式。以上方法均已在项目中进行应用，并取得了良好的效果。

关键词：文档生成；写入生成；自定义模板；信息系统
中图分类号：TP311　　**文献标识码**：A

Research and Application of Document Generation Technology

SHAO Xinxin,ZHANG Minghui,GAO Zijun
(*Department of Software Engineering,Dalian Neusoft University of Information,Dalian* 116023,*China*)

Abstract: Text online generated have a wide range of applications in current information systems.Aiming at the problem of document generation in POI and Itext technology,the difficulties of generating documents are studied,including the generation of the complicated form,the cell's merger etc.Replacing content of Word and Excel is studied.The optimal generation of Word and Excel document is proposed.These methods have been applied in the project,and achieved good results.

Keywords:document generation;written generation;custom template;information system

1 引言(Introduction)

在高校和企业中，报表和文档的处理一直是必不可少的组成部分。当前文档多以电子形式编写和存储，但是很多时候又要纸质版的存档，因此文档的在线生成是一项必不可少的功能。例如电子商务网站的账单、交易额、发票、在线合同等，高校的各类办公和教学文档，例如对于教师有培养方案、大纲、教学日历、教学总结等，对于学生有实验报告、毕设指导手册、开题报告、译文、毕业论文、毕业成绩单、学位证明等重要的文档。这些电子文档往往需要复杂的形式，既包含文本和表格，又包含图片，甚至在表格里加入图片。基于这样的需求，迫切需要一套能够简单、实用、高效地满足各类报表生成的API。本文就以高校的各类文档的生成为需求，展开研究文档生成技术。

当前，常用的Java系列的文档生成的扩展包主要有Itext、POI和JXLS，这几种扩展包可以实现Excel、Word和PDF文件的导出，但是它们都存在某些弊端。Word POI生成简单的Excel的确很优秀，但是操作Word的功能却不尽人意。Itext对于PDF的输出的介绍资料较多，对于Word文档的输出的介绍也不多，对于复杂表格的输出也存在不灵活等问题，而且也无法实现对Word文档分栏和增加水印等功能[1]。和POI结合应用的JXLS在使用模板生成Excel文档方面有一定的优势，而这方面的文献并不多。

综上所述，迫切需要一组API能够实现复杂文档的输出，这也正是本文要解决的问题。本文提出了一套解决方案，具有实用性强、灵活性好等特点。

2 系统框架(System architecture)

高校的文档包含内容较多，如何组织数据和确定数据最合适的输出方式，都是在需求阶段就需要解决的问题，基于文档较多，数据量大的问题，首先对文档进行归类，确定生成方式。本文要研究的文档主要有Word和Excel两种输出形式，根据内容可采用直接生成和模板替换两种方式。

高校文档生成系统共分为四层，包括数据采集层、数据存储层、数据分析和输出层。不同层级间采用文件服务的方式传递数据。系统应用当前比较成熟的SSH框架，前台使用JQuery和AngularJS，数据库采用Oracle数据库[2]，详细的系统架构如图1所示。

基金项目：大连市青年科技之星培育计划项目(ZX2015KJ023).

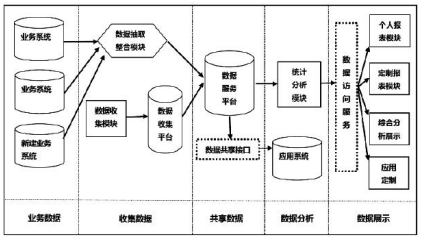

图1 系统架构图

Fig.1 System architecture diagram

3 功能实现(Function realization)

3.1 写入生成

基于对Word POI和Itext现有功能的研究，确定对Itext进行改进后生成Word文档更适合。对现有API进行封装，生成两种方案各自的扩展包，保证具备生成复杂Word和PDF文档的方式和方法。当前的API已经能够实现简单的Word文档的输出，但是对于复杂表格的生成还无法实现。另外，对于生成页眉、页脚和水印等内容也不够灵活。因此，对现有API进行封装，能够实现复杂表格的生成。培养方案、教学大纲和教学总结都采用此方式实现。

项目采用对现有Itext的API进行封装，封装为”ComplexTextUtils”，此类中提供生成word文档的基本方法，包括生成正文、简单表格、复杂表格、水印和页眉页脚等。

这部分的难点就是生成复杂的表格和生成页眉、页脚等内容。

3.1.1 复杂表格的生成

Itext自带的API根据表格内容、行跨度和列跨度三个参数进行组织数据，那么只要计算出行和列的跨度，就可以动态组织数据，生成复杂的表格[3]。

经过改进的API中有多个重载的InsertComplexTable方法，用来实现复杂表格的显示，需要为此方法提供至少三个参数，分别为标题数组、正文数组和列数，如果还有更高的要求，可以提供更多的参数。

public void insertComplexTable(List<CellDetails> titleData,List<CellDetails>contentData,int column,int[] columnwidth) throws DocumentException

那么，如何通过算法对数据的行跨度和列跨度进行计算，是需要解决的重点问题。表格数据多存放于集合中，需要对集合类的结构进行遍历，确定行跨度和列跨度，遍历采用如下方法进行，考虑到章节和知识点之间的嵌套，采用循环遍历的方式对数据进行清洗，确定每个单元和节的行跨度。在当前系统中集合的结构共有三层，相当于树的结构是三层，对内容进行遍历，根据下一层节点的个数，先把上一层结构补充完成[4]。

```
if (skills.size()>1) {
for (Cotlunitskill skill:skills) {
List<Cotlunitsection>sections=skill.getSections();
if (sections.size()<1) {
sections.add(0,createSection(skill.getUnitno()));
} else {
for (Cotlunitsection section:sections) {
List<Cotlunitknowledge>knowledges=section.getKnowledges();
if (knowledges.size()<1) {
knowledges.add(0,createKnowledge(section.getSectionno()));
}
}
}
}
```

通过以上方法，复杂表格的表头如图2所示。

单元标题	节标题	掌握程度	重点	难点	各教学环节学时分配									
					理论授课					实践教学			课内小计	课外小计
					讲课	习题	测验	其他	课外	随堂	实验室	课外		

图2 待生成的表头

Fig.2 Header to be generated

添加表内容后的输出结果如图3所示。

单元标题	节标题	知识点或技能点	掌握程度	重点	难点	能力指标	各教学环节学时分配									
							理论授课					实践教学			课内小计	课外小计
							讲课	习题	测验	其他	课外	随堂	实验室	课外		
Java概述	Java概述	Java特点	1.识记（Remember）			1.2.1	2.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	4.0	0.0
		JDK用法	2.理解（Understand）			8.3.2										
Java基础	Java基础知识	数据类型和变量	3.运用（Apply）	是		1.2.1	10.0	0.0	0.0	0.0	0.0	10.0	0.0	0.0	4.0	20.0	4.0
		运算符	3.运用（Apply）			2.1.2										
	控制语句	条件语句	3.运用（Apply）	是		2.3.1										
		循环语句	3.运用（Apply）	是	是	8.3.2 8.5.3										

图3 生成后的表格

Fig.3 Generated table

3.1.2 页眉、页脚的生成

Itext还能够实现水印、页码、复杂表格、图片、页眉等内容的生成，页眉和页脚生成效果如图4所示。

(a)页眉 (b)页脚

图4 生成的页眉和页脚

Fig.4 Generated header and footer

3.2 替换生成

在高校管理系统中，有部分教学文档格式复杂，但是内容固定，仅需要替换和填充就可以。此类文档适合采用替换的方式生成。教学日历和毕业证明等文档的格式、样式和版式都是相同的，只是部分字段进行替换即可[5]，那么对于这样的文档适合使用文本替换的技术实现，替换涉及的文档主要也是Word和Excel。

3.2.1 Word的替换应用POI的替换功能

首先来看Word文档的替换，主要应用于生成毕业证明和

学位证明。利用标签技术来确定待替换的变量的位置[4]，对于每个要替换的数据都设置为单独的标签，在编写标签需要注意的一点就是标签不能有其他格式，必须清除无效字符，可以在文本编辑器中编写，粘贴到Word文档中采用POI自带的功能，把要被替换的数据组织在HashMap<String,Object>中，例如：

```
WordUtil.inputStream2ByteArray(new FileInputStream("icon.jpg"), true));
param.put("${header}", header);
```

可以替换文字、日期和图片等内容。

通过下面的语句获取文档的段落和内容，进行替换。

```
List<XWPFParagraph>paragraphList=doc.getParagraphs();
List<XWPFRun>runs=paragraph.getRuns();
Iterator<XWPFTable>t=doc.getTablesIterator();
while (it.hasNext()) {
    XWPFTable table=t.next();
    List<XWPFTableRow> rows=table.getRows();
对行内容进行处理；}
```

替换的模板如图5所示。

姓名：$STUDENTNAME$，性别：SEX，$BIRTHDAYSTR$生。该生于$ENTERYEAR$年九月至$GRADYEAR$年七月在[redacted]$SPECIALITYNAME$专业$LEVELNAME$层次学习，学制$DURATIONYEAR$年。该生已获得$DEGREESUBJECT$学士学位证书（证书编号：$DEGREECERTNO$）。

图5 待替换的word模板

Fig.5 Word template to be substituted

替换后的结果如图6所示。

姓名：张小小，性别：女，1995年12月生。该生于二零一一年九月至二零一五年七月在[redacted]软件工程专业本科层次学习，学制四年。该生已获得工科学士学位证书（证书编号：1111111111）。

图6 替换后文本

Fig.6 Substituted text

3.2.2 Excel的替换

其次是Excel的替换。Excel文档的替换主要是在教学日历中使用，经对比发现利用JXLS来实现内容的替换最适合。此方法采用编写脚本的方式，脚本中可以写入代码，相同格式的数据可以循环插入，把数据以类的方式组织，放在List中，例如Course类。如果Course类中包含其他的类，可以作为属性[7]。

```
List<Course>course=createCourse();
Map beans=new HashMap();
beans.put("course",course);
```

模板内代码如下所示。

```
<jx:forEach
 items="${courses}"var="course"></jx:forEach>
```

总结出以下几点：(1)模板替换的好处是组织的实体类中某个属性的创建，值为null时，无需处理数据。(2)JXLS这套API对图形和图表的支持很有限，而且仅仅识别PNG格式。(3)Excel的替换支持多sheet页的替换，具有一定的可行性。

如果数据是多条，采用下面的语句数据会重复显示。例如，下面是五条数据，显示结果如图7所示。

```
List<Check> checks=new ArrayList<Check>();
checks=Check.generateCheck(5);
```

考核分类	考核项目	满分值
items="${checks}" var="check">		
${check.category}	${check.item}	${check.score}
</jx:forEach>		
总计		${SUM(C4)}

考核分类	考核项目	满分值
形成性考核	考勤	5
	考勤	5
	考勤	5
	考勤	5
	考勤	5
总计		25

图7 模板替换生成文档效果图

Fig.7 Document sample of template substitution

4 结论(Conclusion)

通过对Itext、POI和JXLS等文本输出技术的研究和对比，对原有API进行改进，采用最优的方法实现了文本的在线输出，能够满足以多种复杂的方式输出Word和Excel文本。本文的方式和方法已在项目中进行应用。另外在文献[6]—文献[8]的App系统中，也进行了应用，并取得了非常良好的效果。

参考文献 (References)

[1] Bing L,Li P,Liao Y,et al.Abstractive Multi-Document Summarization via Phrase Selection and Merging[J].Computational Linguistics,2015,31(4):505-530.

[2] Abualigah L M,Khader A T,Al-Betar M A,et al.Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering[J].Expert Systems with Applications,2017,84(C):24-36.

[3] Ayyar K,Han C J.System and method for computing, applying, and displaying document deltas[J].Remote Sensing of Environment,2017,118(118):339-355.

[4] 周千明,朱欣娟.基于Aspose技术的自定义模板文档生成方法[J].计算机系统应用,2015,24(6):235-238.

[5] 张艳伟.QT框架下的WORD文档生成方法[J].计算机应用与软件,2015,32(10):120-122;150.

[6] 付丽梅,邓继禹,贾跃.基于腾讯微校平台的易学习APP设计与实现[J].考试周刊,2017(07):112.

[7] 付丽梅,刘英鹏,贾跃.基于腾讯微校平台的校园移动办公APP设计与实现[J].信息系统工程,2017(01):156-157.

[8] 邵欣欣,徐健晟,冀航宇.基于VR的幻房App的设计与运营分析[J].电子元器件与信息技术,2017,7(01):1-5.

作者简介：

邵欣欣(1980-)，女，硕士，副教授.研究领域：软件工程，虚拟现实.

张明会(1980-)，女，硕士，教授.研究领域：软件工程，算法.

高梓峻(1995-)，男，本科，工程师.研究领域：软件工程，软件开发.