基于 OpenXML 的论文格式检测系统的设计

王艺焱[1] 林刚[1(通迅作者)] 陈小冰[2] 黄逸轩[1] 袁笛[1]

1 珠海科技学院计算机学院，广东珠海，519041；

2 新南威尔士大学，澳大利亚新南威尔士州悉尼肯辛顿，2052；

摘要：针对人工检测论文格式存在效率低下、错漏率高、反馈周期长及错误定位不便捷等问题，设计了一款基于 OpenXML 的论文检测系统。该系统以本科毕业论文格式为核心，依据国家及珠海科技学院本科毕业论文格式规范，提出格式信息解析方法，设计了模板创建、格式识别、论文检测三个基础功能模块。其中，模板创建模块负责解析格式模板并生成模板对象；格式识别模块负责提取上传论文的格式信息并输出待检对象；论文检测模块通过比对模板对象与待检对象，自动生成格式检测报告。测试表明本系统能显著提升格式检测的效率与管理效能，同时具备良好的便捷性、灵活性与可扩展性。

关键词：论文格式检测；面向对象；Docx4j；Open XML；Java

DOI：10.69979/3041-0673.26.04.003

引言

各高等院校在学位论文的内容上与格式上都有严格的规定[1]，论文质量是衡量高校教学水平的关键指标，也是评价毕业生创新人格、思维与技能的重要依据。然而，目前本科毕业论文普遍存在质量不高的问题。杨月元在研究本科毕业论文写作规范[2]中提及，格式不规范是导致该问题的重要原因之一。学生对论文重视轻、投入少以及粗阅研读格式规范等，都会导致论文格式错误，进而加重教师批阅负担。同时部分导师在指导中也存在重形式、轻内容的问题[3]，进一步影响了论文质量。

在姚世斌等人制定的本科毕业论文质量评价体系中指出“格式规范”占论文综合质量的 26.7%[4]，所以提高论文格式规范程度，在一定程度上能提高本科毕业论文的质量。随着本科毕业论文格式要求的日益严格，高效、自动化的格式检测解决方案变得尤为迫切，毕业论文格式检测系统的研发已成为学术界关注的重点，研究者们相继提出了流式文档排版自动化测试方法[5]、论文格式检测系统[6]以及高校毕业论文格式设置软件[7]等。为满足日益严格的格式检测需求，本文研究了一种基于模式识别的毕业论文格式检测方法，并基于此方法设计了一款 Web 系统。该系统使用 Open XML 技术提取文档信息以执行自动化的格式规范检查。

1 需求分析

本系统旨在为高校提供一种高效便捷的论文格式检测服务，主要服务对象是老师和学生。从用户的角度来说，论文格式检测系统需要提供四个基本功能：

a) 模板创建：支持用户根据需求创建或上传标准模板对象，实现个性化定制。

b) 格式识别：基于预设模板识别用户上传的论文文档，形成结构化的待检测对象。

c) 论文检测：系统通过将待检测论文对象与选定的标准模板对象进行智能比对，确保论文格式严格遵循学术规范。

d) 检测报告：系统生成结构清晰、内容详尽的格式检测报告，帮助用户快速定位问题并进行针对性修正。

根据以上系统为用户提供的功能，在设计论文格式检测系统时需要包括模板创建、格式识别、论文检测、检测报告等功能，这些功能都要遵循一定的格式标准。在模板创建阶段，系统将依据国家标准 GB/T 7713.1-2006 学位论文编写规则[8]定义的模板文件或者本科院校自定义的论文模板文件进行解析，生成标准模板对象，解析的内容包括论文前置部分、正文部分、参考文献、附录以及结尾部分[9]。格式识别阶段要求系统对学生上传的论文检测稿文件进行格式识别，通过解析其文档结构、样式信息及内容元素，生成待检测对象。论文检测阶段要求系统将待检测对象与标准模板对象进行比对，检测的内容包括封面、章节、目录、图、表等样式检测。最终，系统将基于比对结果生成一份详细的格式检测报告，明确指出存在的格式偏差及其具体位置。

2 功能设计

针对上述提及的论文检测功能模块，本章分析论文检测时的功能需求，提出待解决的问题。

a)对于模板创建部分，需要从论文模板中提取文本和格式信息用于检测格式错误。在创建本科毕业论文模板时，不同地区、不同学校、不同学院对论文格式的细节要求均不相同。但本质上都是由如下部分组成：封面、目录、中英文摘要、关键词、正文、注释（脚注或尾注）、参考文献、附录、致谢等。可见论文组成部分复杂且格式要求多，所以需要从论文模板中提取重要的格式条目，该格式条目应该具备论文不同组成部分的特定要求，比如在摘要页上，关键字的位置与格式，摘要题目的摆放位置字体大小等都是摘要页上特有的格式条目。将此类格式信息整合后生成标准模板对象，包含所有标准模板格式类别、格式名称和格式内容。

b)对于格式识别部分，需要对学生论文检测稿文件格式进行识别。根据选定的标准模板对象的格式条目对论文进行扫描提取，生成待检测对象；也就是说待检测对象的待检测格式条目需要与标准模板对象的相应条目对应。

c)对于论文检测部分，需要将待检测对象与标准模板对象进行比对时，包括格式条目匹配与标号匹配。在进行格式条目匹配时，待检测对象需要与标准模板对象中的条目一一对应，然后我们将两个对象的相对应条目的特定格式进行比对，比如标准模板对象的摘要标题字体对应的特定格式是“宋体 三号”，那么相对应待检测对象摘要标题字体的格式也应该是“宋体 三号”，否则匹配失败，或格式错误。标号匹配即对整篇论文中的章、节和图、表标号等进行匹配，该检测功能要保证标号的连续性和正确性。

2.1 模板创建

模板创建功能需要从论文模板中提取文本和格式信息用于格式比较。论文的主要组成部分对格式的要求均比较多，为了便于检测，需要为每个部分做进一步细分，提取重要的格式条目。格式条目应该反映论文相应组成部分的特定要求，比如在摘要页上，摘要标题位置、字体大小、关键字的位置与格式等都是摘要页上特有的格式条目。所以格式条目需要包含格式类别、格式名称、和格式内容三个属性。格式类别即该格式条目所属的论文部分；格式名称即扫描标准文档后提取出来的待检测格式的属性名称；格式内容即待检测格式的属性值。将此类格式信息整合后生成标准模板对象，包含所有标准模板的格式类别，格式名称与格式内容。

本文结合上述提到的国家标准 GB/T 7713.1-2006 学位论文编写规则所总结出的本科毕业论文主要组成部分和珠海科技学院毕业论文格式要求，对论文的格式条目包含的三个属性做出说明。

a)格式类别，包含模块，构成元素，格式类别编号。比如列出“封面论文标题”的格式类别时，“模块”是封面；“构成元素”可以是“题名”，“基本信息”等元素名称；“格式类别编号”即该构成元素的唯一编号。

b)格式名称，即论文构成元素的格式要求，比如“字体”、“字号”等。

c)格式内容，即论文构成元素的格式要求对应的属性值，其值与格式名称一一对应，比如格式名称“字体”对应格式内容“宋体”。其值可随论文模板的细微修订而变化。

2.2 格式识别

与模板创建过程相类似，格式识别阶段同样需要对论文文档进行结构化解析，其目的在于生成可供系统处理的待检测对象。然而，二者在解析目标与应用范围上存在本质区别：模板创建所处理的是符合规范的标准模板文件，通过解析生成具有通用性的标准模板对象，该对象可被所有采用同一格式标准的论文检测任务重复使用；而格式识别则面向学生提交的个体论文文件，通过解析生成仅针对当前检测论文的个体待检测对象，其有效范围限定于单次检测过程。

若在检测过程中采用简单的全局比对策略，直接将模板与整篇论文进行一对一的匹配，系统将面临处理效率低下、灵活性不足等突出问题，在实际应用场景中缺乏可行性。为此，本系统引入了分块检测机制，将整篇论文按照逻辑结构划分为多个相对独立的单元，并为每个单元配置针对性的检测规则。这种模块化的处理方式不仅显著提升了系统的处理效率，还能够根据不同章节的格式特征实现更精细化的检测，从而有效提高格式识别的准确率。

基于国家标准 GB/T 7713.1-2006《学位论文编写规

则》的要求，并结合珠海科技学院本科毕业论文格式规范的具体实施需求，本系统将学位论文的结构划分为三个主要部分：前置部分、主体部分和结尾部分。为简化系统架构并增强核心模块的复用性，将参考文献纳入主体部分进行统一管理。通过这种层次化的结构划分，系统能够更有针对性地实施格式检测，既保证了检测过程的全面性，又提高了系统的可维护性和扩展性。主体部分内容为正文、图表以及参考文献。要说明的是本科毕业论文封面部分不包含页脚；剩余部分包含两种页脚，在摘要部分和目录部分使用罗马数字，正文以下的部分则是阿拉伯数字。

2.3 论文检测

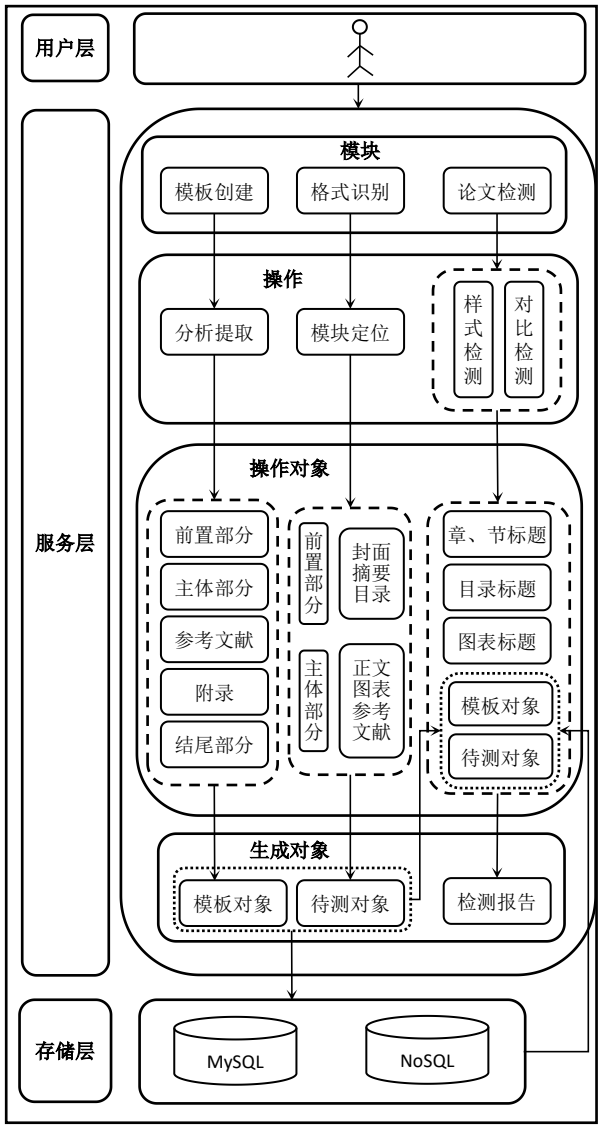
用户上传论文后，系统将自动选择模板进行全面检测包括封面格式、章节标题、目录结构、图表标注、参考文献等著录要素在内的各类格式规范，覆盖常见的字体、间距、标号、引用样式等细节问题，检测完成后生成报告。在检测时首先需要对每个模块进行定位操作，定位的方式通过查找关键词等方法实现。当系统定位到对应模块时，就可以在预先存取的模板格式数据中提取此模块的格式数据并使用它来对论文进行检测，且提取出来的格式数据与检测论文时某一格式要求一一对应。系统架构与实现。

3 系统架构与实现

本系统基于 Open XML[10]标准，使用 Docx4j 解析 docx 文档结构，以进行精准的自动化格式检测。平台采用 SpringBoot+Vue.js 前后端分离架构，结合 MySQL 与 MyBatis 进行数据处理，并通过 Dom4J 确保论文格式符合 OOXML 规范。

3.1 系统结构

系统支持用户将论文批量上传，选择某一篇论文进行检测以及错误报告导出功能。系统还支持教师定制论文模板，教师可以针对七个模块进行自主选择，选择内容由系统给出并支持模板删除功能。在论文检测时可进行模板选择功能，选择的模板由教师创建而来，也可选择系统默认模板。支持错误报告导出，在论文检测完后会生成一篇错误报告交予用户阅读并自行修订。支持用户登录、注册功能，用户可通过账号密码使用系统。系统整体框架图如图 1 所示。



3.2 系统测试

为了全面评估系统对论文格式错误的识别能力，本研究设计了两类测试用例。第一类采用符合规范的《珠海科技学院计算机科学与技术专业毕业论文模板》作为标准样本，用于验证系统的正确识别能力。第二类选取 49 份包含常见格式问题的论文稿件作为测试样本，用于检验系统在不同类型格式错误场景下的检测准确率与鲁棒性。测试涵盖了格式规范的各个方面，包括但不限于字体、间距、页眉页脚、参考文献格式及图表标注等关键要素，以确保测试的全面性和系统性。

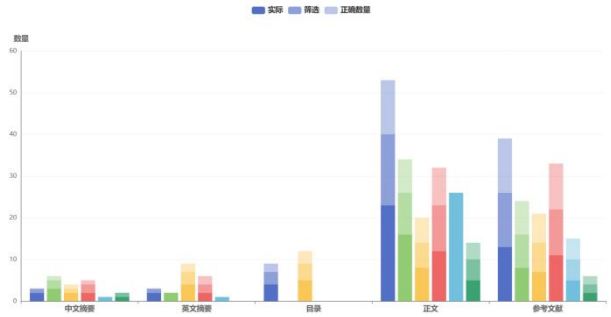
评估系统准确率时，我们将统计正确识别的错误格式数量、实际识别出的错误格式总数量以及筛选后的格式错误总数量。利用这三个数据计算出错误格式识别率。其中，正确识别的错误格式数量是系统检测到的并且与实际错误相符的数量。实际识别的错误格式总数量是测

试文档中所有的格式错误批注数量。错误格式识别率计算公式如下：

$$F=\frac{C}{A}\times 100\% \quad (1)$$

式中 F 为错误格式识别率，C 为正确识别的错误格式数量，A 为实际识别的错误格式总数量或者筛选后的格式错误总数量。

第一类论文格式准确度测试部分结果如图 2 所示。



从整体识别效果来看，系统在全部 413 个实际错误中成功筛选出 311 项疑似问题，并最终准确识别出其中的 271 项，整体错误识别率为 65.6%。

分析发现，由于论文内部的 XML 文件较为复杂，系统在检测时经常会遇到误判问题。例如：

a)系统可能因为文件格式的问题无法检测到目录，导致目录部分出现大量错误。

b)系统可能无法检测到带有超链接的论文标题，进而影响标题检测和顺序检测。

c)系统可能无法正确处理包含特殊字符（如希腊字母、数学符号等）的段落。这些特殊字符可能会导致系统误判为格式错误，影响检测结果的准确性。

在第二类测试中，系统对 49 份论文稿件的格式错误进行了全面检测。测试样本根据质量等级划分为四类，包括 15 份优秀论文、15 份良好论文、15 份中等论文以及 4 份及格论文。为深入分析系统性能，本研究对每类论文分别统计了实际错误数量、筛选后错误数量及系统正确识别的错误数量。每篇论文进一步划分为六个结构模块，分别记录各模块中实际存在的错误数量与系统识别正确的错误数量。论文错误统计结果见表 1。

论文种类	实际识别错误格式总数量	筛选后的错误格式总数量	正确识别的错误格式数量	筛选后的错误格式识别率	实际的错误格式识别率
优秀论文	710	501	439	87.62%	70.56%
良好论文	201	130	123	94.62%	64.68%
中等论文	413	311	271	87.14%	75.30%
及格论文	520	379	333	87.86%	72.88%

测试结果表明，在存在系统误判的情况下，错误格式识别率处于 64%至 73%的区间；而经过人工筛选排除误判后，系统的识别率可提升至 87%至 95%之间。具体来看，系统对优秀、中等和及格等级论文的错误格式识别率均超过 87%，对良好论文的识别率更是达到 94%以上。经筛选后的整体错误格式识别率平均值为 89.31%。由此可知，通过筛选后的错误格式总数量来计算错误格式识别率，可以提高检测结果的准确性，减少误判的影响，更好地评估和优化系统性能，并确保研究的严谨性和可信度。

4 结论

本文设计并实现了一个基于 Open XML 的论文格式检测系统。该系统通过定义结构化的格式属性，将国家标准与院校规范转化为可计算的数据模型，并利用面向对象方法构建了模板创建、格式识别与论文检测三大核心模块，实现了自动化格式比对。测试表明，系统在对各类论文稿件的检测中，经筛选后的平均错误格式识别率达到 89.31%，能显著提升格式检测的效率与准确性，有效减轻人工审阅负担。

尽管当前系统在自动化检测方面取得了良好效果，但其对文档内部复杂结构（如非标准目录、超链接等）的适应性仍有提升空间。下一步工作将着重优化文档解析算法，增强系统对特殊格式元素的鲁棒性，并计划将检测范围扩展至更多院校和类型的学术文档，以进一步提升系统的通用性与实用价值。

参考文献

[1]吕秉原,李敏.基于NLP技术的高校毕业论文智能评估系统的设计与实现[J].电脑知识与技术,2025,21(25):41-47+52.

[2]杨月元.浅谈应用型本科毕业论文写作规范问题[J].教育现代化,2018,5(49):240-241.DOI:10.16541/j.cnki.2095-8420.2018.49.079.

[3]王德朋.本科毕业论文的功能应重新定位[J].中国大学教学,2017,No.321(05):49-52.
[4]左阔，李宁，田英爱，等.流式文档排版效果自动化测试方法[J].计算机工程与应用，2021,57(2):273-278.
[5]徐俊.毕业论文格式检测与校正系统研究与实现[D].重庆：重庆邮电大学，2022.
[6]董建文.高校毕业论文格式设置软件的设计与实现[J].智能计算机与应用,2025,15(02):70-76.
[7]姚世斌,彭宇霞,潘艳等.基于学术规范的本科毕业论文质量评价体系建设[J].高教探索,2016(S1):98-99.
[8]中国国家标准化管理委员会.(2006).学位论文编写规则:GB/T 7713.1-2006[S].北京:中国标准出版社.
[9]国家标准委员会.机械工程技术文件编制规则:CY/T 35-2001[S].北京:中国标准出版社,2001.
[10]廖根为,凌治博.OOXML 字处理文档的 RSID 变化规律与鉴定价值分析[J].中国司法鉴定,2025,(03):74-82.

基金项目：1）广东省本科高校教学质量与教学改革工程项目：计算机类《课程设计》类课程的过程化考核研究与实践（2024008）；
2）珠海科技学院产教融合型课程培育建设项目：计算机系统综合课程设计（CJRH2023009）；
3）珠海科技学院高等教育教学改革项目：课程设计过程化考核研究与实践（ZLGC20230706）。