诚 信 声 明

我声明，所呈交的毕业论文是本人在老师指导下进行的研究工作及取得的研究成果。据我查证，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得其他教育机构的学位或证书而使用过的材料。我承诺，论文中的所有内容均真实、可信。

毕业论文作者签名： 签名日期： 年 月 日

基于OOXML的毕业论文检测和校正系统

**[摘 要]**

**[关键词]** 格式检测；格式校正；学位论文；OOXML；XML

Paper detection and correction system based on OOXML standard

**Abstract：**

**Keywords：**Format Detection；Format Correction；Dissertation; OOXML; XML；

目 录

[1 绪论 1](#_Toc16966)

[1.1 研究背景及意义 1](#_Toc23878)

[1.1.1 研究背景 1](#_Toc24001)

[1.1.2 研究意义 2](#_Toc1741)

[1.2 国内外研究现状 2](#_Toc28508)

[1.3 论文主要工作 4](#_Toc8456)

[1.4 论文组织结构 4](#_Toc18092)

[2 相关技术 6](#_Toc28596)

[2.1 文档技术 6](#_Toc2290)

[2.1.1 XML 6](#_Toc6562)

[2.1.2 DOC/DOCX 6](#_Toc17927)

[2.1.3 OOXML 7](#_Toc2666)

[2.2 开发技术 7](#_Toc29114)

[2.2.1 DOM4J 7](#_Toc10960)

[2.2.2 Spring Boot 8](#_Toc19754)

[2.2.3 MySql与MybatisPlus 8](#_Toc27963)

[2.2.4 Vue3与Vite 9](#_Toc26589)

[2.2.5 ElementPlus 9](#_Toc22248)

[3 需求分析 10](#_Toc24228)

[3.1 系统需求概述 10](#_Toc13254)

[3.2 平台交互需求 10](#_Toc28303)

[3.2.1 基础功能 11](#_Toc21018)

[3.2.2 论文提交与论文模版管理功能 12](#_Toc25717)

[3.2.3 批量生成账号与指导关系 13](#_Toc19550)

[3.3 内容无关格式检测 13](#_Toc30182)

[3.3.1 标题格式检测 14](#_Toc23286)

[3.3.2 段落格式检测 15](#_Toc13950)

[3.3.3 其他格式检测 15](#_Toc13488)

[3.4 内容相关格式检测 16](#_Toc13548)

[3.5 生成检测报告与自动校正 17](#_Toc31790)

[4 WordProcessingML文档分析 19](#_Toc25909)

[4.1 文件组织结构 19](#_Toc27046)

[4.2 控制基本样式 21](#_Toc27123)

[4.2.1 基本内容和样式的组织结构 21](#_Toc7070)

[4.2.2 分层级的样式控制 22](#_Toc7325)

[4.2.3 常见样式控制标签 23](#_Toc3724)

[4.3 控制页面尺寸与分节 24](#_Toc29992)

[4.4 控制页眉页脚 25](#_Toc31369)

[4.5 控制编号 25](#_Toc13174)

[4.6 控制图表和文本框 26](#_Toc3954)

[4.7 控制批注 28](#_Toc363)

[5 系统设计 30](#_Toc2854)

[5.1 系统总体结构 30](#_Toc15024)

[5.2 数据库设计 30](#_Toc17105)

[5.3 平台交互功能设计 34](#_Toc12126)

[5.3.1 普通用户端 35](#_Toc16219)

[5.3.2 教师端 37](#_Toc8486)

[5.3.3 管理员端 37](#_Toc20081)

[5.4 论文检测程序设计 38](#_Toc30396)

[5.4.1 内容无关格式检测功能设计 39](#_Toc13512)

[5.4.2 内容相关格式检测功能设计 40](#_Toc16265)

[5.4.3 报告生成与自动校正功能设计 41](#_Toc21385)

[6 系统实现 42](#_Toc12781)

[7 系统测试----- 43](#_Toc25095)

[结论 44](#_Toc24730)

[致谢 45](#_Toc9769)

[附录A 46](#_Toc2007)

[参考文献 47](#_Toc10546)

# 绪论

## 研究背景及意义

### 研究背景

学位论文也是高校学生培养计划的关键环节，既是对学生在大学期间所学知识、学术科研能力和独立思考能力的一次重要考察[1]，是决定是否授予学位的重要评判指标，也是对高校教学能力和科研水平的一次展现。因此，学位论文的质量对学生和高校都至关重要。其中，保证论文质量的一个重要手段便是保证论文的格式规范性[2]。无论是对论文内容的精准表达，保证论文的学术性和严谨性，还是方便学者阅读交流和传播，都是建立在论文的格式规范性之上。

当前撰写论文常用的编辑工具主要有LaTex、Microsoft Office Word和WPS Office三种。LaTex是一种基于TEX的排版系统，功能强大，支持多种复制的排版功能，但相较于后两者而言LaTex的学习成本更高，对于无计算机科学基础的人来说，需要耗费大量的时间去学习“环境变量”等知识， 并且由于LaTex并非“所见即所得”，需要编译才能看到效果，初学者往往容易遇到编译出错的情况且难以排查。因此，国内高校学生大多都使用后两者编辑工具来撰写论文，文档格式上都是使用Word文档格式，即.doc文件和.docx文件。学校设计的学位论文模板也大多是采用Word文档格式作为模板。

学生在实际撰写学位论文的过程中，即使学校已经提供了论文模板文件来辅助论文排版，学生在将论文送审或参与毕业答辩时，无论论文内容如何，学生水平高或低，都会出现不同程度的格式排版问题[3]。这一方面是因为学位论文的规范要求比较繁杂，大到页面设置、图表注释、封面、页眉页脚、摘要、目录、绪论、正文、结论、参考文献、致谢等大块内容，小到中英文字体类别、字体大小、段间距、段前距、段后距、对齐方式、引用格式等细致内容，学生完全理解和掌握这些要求并不容易；另一方面，许多学生本身对使用Word软件不熟练，在调整间距、绘制图标等功能上非常生疏。加之当前绝大数高校并未给学生提供自动化的格式检测工具，这客观上导致了对各种细枝末节的格式问题只能靠学生和导师手工调整，无可避免地占用了学生和导师大量的时间和精力，尤其是在学生素质整体较低的高校中，这种时间和精力的占用问题更为明显。一份来自常州信息职业学院的调研[4]指出，有63%的学生因格式问题重复修改论文达到3次以上，近80%的学生认为调整论文的格式占据了整个论文撰写30%以上的工作量。此外，格式处理分散了学生和导师的精力，也不利于高质量的学位论文的产出。

自XML技术出现以来，包括微软在内的公司逐渐意识到在产品中使用开放格式和标准规范的重要性和商业价值。直到Microsoft Office 2007系列发布，Word文档形成了以XML模式为基准的ECMA-376文件标准，即OOXML[5，6]（Open Office Extensible Markup Language）文件标准，该标准对Word文档如何进行文件组织、压缩打包、内容样式控制等作了详细的规定。至此，Word文档抛弃了过去以二进制格式存储的.doc文件，改用了以XML文件为底层的.docx文件，这不仅意味着Word有了公开的、通用的文件标准，也使得开发自动化计算机程序对使用Word文档撰写的论文进行自动化格式检测和修订成为了可能。

### 研究意义

本文研究与实现基于OOXML的毕业论文检测和校正系统的意义主要有如下三点：

1. 极大程度减轻人工调整论文格式的负担，在撰写论文时回归“重内容，轻格式”，留出更多精力在论文内容上，提高论文质量。对于学生而言，学生可以省下许多扣细节格式的时间，也不必学习太多Word文档格式调整功能，就能轻松提高学位论文的规范性；对导师而言，可以不必担心学生论文的格式问题，留出更多精力去给学生论文在撰写方向、内容逻辑、用词表达方面的指导，提高指导效率和学位论文的质量。
2. 深入研究文档底层的OOXML结构，充实了Word文档自动化处理领域的相关知识，并拓展了格式自动化处理的方法论。这些研究成果为企业在进行文档自动化开发时提供了宝贵的参考。

3. 构建的一种对毕业论文的检测和自动校正的系统，能给学校、企业提供一种可行的、可借鉴的论文格式检测方案，具有实用意义。

## 国内外研究现状

国外对于学位论文检测的相关内容更多的是集中在抄袭检验，防止学术作弊上，侧重对论文内容进行分析，在格式检测方面的研究比较少。此外，由于语言文字上的不同，国外学位论文格式规范要求和国内的规范要求差异很大，所以分析针对国内对学位论文的研究相比针对国外学位论文的研究会更有针对性。

当前针对Word文档的格式检测大体上有两类技术方案。一是依托Word对象模型处理文档，例如北京邮电大学的张印于2012年设计并实现的基于VSTO的论文格式处理软件[7]，大连理工大学的张巍巍于2019年实现的论文格式规范性自检系统[10]。另一种是依托OOXML标准处理Word文档的底层XML文件，例如大连理工大学的陈渊博于2017年实现的学位论文格式检测系统[8]，重庆邮电大学的徐俊于2022年实现的毕业论文格式检测系统[9]，北京交通大学的袁敏于2019年关于学术论文格式检查和内容校对研究[11]。

依托Word对象模型处理文档的本质是通过程序间接操作Word编辑软件，要求服务器上提前安装好Word编辑软件，对服务器的操作系统环境要求较高，又因为其本质是间接操作，执行效率不高。

张印的系统[7]使用.NET Framework 框架开发，通过程序调用库函数，操控当前系统上的 Microsoft Word 组件。系统主要提供对论文标题、参考文献、图形、表格部分的校验，检测项目比较少，最后以单独输出的检验报告形式反馈用户，用户根据报告内容修改原文件，没有自动修改功能。

张巍巍的系统[10]在ASP.NET开发框架下，使用Word对象模型提供的主互操作集与COM对象模型进行交互，从而实现用程序控制Word文档。相对于张印的系统，张巍巍系统的检测项更完善，且采用一种模块识别与定位的方法，先确定是否有部分缺失，再进行检测，提高了检测效率。生成的用户报告提供PDF版本，但也是没有自动修改功能。

依托OOXML标准处理Word文档的方式不需要提前安装Word软件，本质是处理Word文档解压后的XML文件，效率更高，但难点就在于分析Word底层的组织模式，且无法直接操作doc文件。

徐俊的系统[9]使用开源框架技术Apache POI操作Word底层XML文件，实现了对系统的校正和检测，还实现了对内容上一些标点符号的检测，但是对参考文献的检测比较薄弱，也不支持doc文件格式。

陈渊博的系统[8]基于Open XML SDK对论文进行检测，并用SSH框架开发出了提供检测服务的Web网站。系统性能好，能覆盖大部分格式问题，但是提供的检测报告是txt格式，且不提供自动修正功能，学生需要一条条对照自己改进，在论文篇幅较大时对照修改并不方便。系统进行检测时也不支持doc文件。

袁敏的系统[11]也是使用开源的Apache POI操作Word底层XML文件，能实现对论文段落的字号、字体、缩进等常规文本规范的检测，但相比之下系统更侧重对论文内容校对上，例如词语语法、错别字的检测。系统最后也是提供一个检测报告，需要用户自己对照修改，无法实现自动修正。

总结当前的论文格式检测系统，各有优势也各有不足，可以总结得到论文检测系统当前需要改进的问题如下：

1. 保证检测范围的情况下尽可能提高检测效率。
2. 提供自动修正功能，无需用户自己手改。
3. 提供易对照的检测报告，方便用户对照修改。
4. 提供更好的格式兼容性，使系统兼顾doc和docx文档，并提供输出PDF文档的功能。

## 论文主要工作

本文的主要工作在于针对现有论文格式检测系统的不足，设计并实现了一种基于OOXML标准处理Word文档论文的格式检测系统。文本的主要工作如下：

1. 根据OOXML标准分析Word文档底层的XML文件组织方式和内容组织方式。微软公司Office2007版本及其之后版本都是依据OOXML标准组织文件底层，分析理解Word文档底层的XML文件组织方式和内容组织方式，才能进行一步使用程序操作Word文件。论文详细分析了Word文档底层是如何用XML标签控制页面的样式、页眉、页尾、编号、图片表格、图注表注、字体样式等内容的显示和变化方式，分析各种控制标签的含义。
2. 实现一个依托OOXML标准进行Word文档格式检测的程序。基于OOXML标准处理Word文档，无需提前安装Word软件，能具有更高的处理效率。在程序中对传入的Word文档进行格式转换并解压为XML文件，随后对XML文件进行处理，对论文的各个部分进行定位，分别检测格式，并实现对部分格式的自动修改。在检测范围上，覆盖所有常见的论文格式问题，包括字体、字号、行距、缩进、页边距等常规文本规范，以及标题、段落、列表、表格、图片等文档结构的规范性。另外，系统内部会对传入的doc文档进行格式转换，将其转换为docx格式，提高文档的兼容性，并对于检测通过的论文能自动导出PDF格式。
3. 实现以批注形式展现的检测报告生成程序。在检测过程中，系统会对每个发现的问题进行标注，并以批注的形式直接显示在Word文档中。这样，用户最终得到的检测报告是以批注形式展现的，而非一个单独的检测报告，用户可以直接在Word文档中看到所有需要修改的地方，无需再对照其他格式的检测报告，使修改过程更加直观和便捷，用户只需点击批注，即可直接定位到问题所在，大大提高了修改效率。
4. 实现了一个毕业论文格式检测系统。基于论文检测程序与报告生成程序，用vue实现了系统的前端，用springboot和mybatisplus实现了系统的后端，形成了一个完整的检测系统。Web前端提供用户端、教师端和管理员端，能满足高校中不同角色使用论文格式检测系统的不同需求。

## 论文组织结构

论文全文共七个章节，围绕Word学位论文格式检测与修订系统的实现和改进需要进行研究，各章节内容主要安排如下：

第一章为绪论，介绍本课题的研究背景、研究意义和相关研究现状，分析了实现Word论文格式检测的主要技术实现方案，列出了本文主要的研究内容和针对Word论文格式检测需要改进的问题，并给出论文组织结构。

第二章为相关技术介绍，主要介绍了本系统用到的文档技术和开发技术。

第三章为需求分析，从格式规范检测需求、内容逻辑检测需求和平台交互需求三个方面分析系统的需求。

第四章为OOXML文档分析，介绍和分析了Word文档底层如何用XML文件组织内容。

第五章为系统设计与实现，主要介绍了系统功能设计、数据库设计以及各个用户端的具体功能设计和实现。

第六章为系统测试，介绍了对系统检测功能的测试方案和测试结果。

最后是对本系统研究工作的总结，并对系统未来的进一步完善和改进进行了展望。

# 相关技术

## 文档技术

### XML

XML，全称是可扩展标记语言（Extensible Markup Language），它是一种用于编码文档的标记语言。可扩展标记语言（XML）提供了一种标准化的方式，使得数据能够以一种可共享和通用的格式进行定义和存储。它促进了不同计算机系统（如网站、数据库和第三方应用程序）之间的信息交流和协作。XML 定义了严格的规则，这些规则极大地简化了数据以 XML 文件形式在网络上的传输过程，确保了数据的准确性和高效性，使得接收者能够按照这些规则轻松解析和获取所需的信息。

不同于HTML，XML被设计用来传输和存储数据，而非显示数据。HTML是为显示数据而设计的，而XML则是用来描述数据的。

XML主要具有以下特点：

1. 自描述性：XML文档自身包含了文档的结构或数据的描述信息。这一特性使得软件能够处理具有不同结构的数据，而无需事先了解这些结构的详细信息。
2. 简洁性：XML的设计理念在于其易读性和易写性。同时，XML文档也易于在计算机系统中进行处理和分析。
3. 开放性：XML基于开放的标准，独立于特定的软件供应商。无需注册专利，它是免费可用的，并且适用于各种目的和应用场景。
4. 可扩展性：XML允许用户定义新的元素和属性名称，以适应特定应用的需求。这一特性使得XML文档能够根据实际的应用场景进行定制，以满足特定的数据表示和交换要求。

### DOC/DOCX

DOC文件是Word 2003及之前版本的主要文档格式，是微软的一种专有格式，最初主要用于纯文本文件，常见于不同的操作系统中，以及软硬件的使用说明。DOC文件是二进制文件，能容纳许多文字格式、图片、[脚本语言](https://zh.wikipedia.org/wiki/%E8%85%B3%E6%9C%AC%E8%AA%9E%E8%A8%80" \o "脚本语言)及撤销等信息，在功能上比其他的文件档格式如RTF、HTML等更为丰富。然而，由于它是封闭的格式，其兼容性相对较低。

DOCX文件是Word 2007及以后版本的主要文档格式，它是基于Microsoft开发的OOXML（Office Open XML）规范组织的文件，底层使用的是XML语言文件格式，因此具有可开放性和可拓展性。与DOC格式相比，DOCX在多个方面都有所改进。例如，DOCX的文件存储更高效，文件大小通常比DOC文件小很多，这主要得益于其采用的XML格式能更好地压缩文件大小。此外，DOCX文件的内容更为丰富，支持更多的文本样式、图像、表格和多媒体内容。在文件结构方面，DOCX更为灵活，能更好地支持大型和复杂的文档结构，同时方便与其他文件格式进行转换。最后，DOCX文件的安全性也更高，支持数字签名和加密技术，能更好地保证文档的安全性和完整性。

### OOXML

Office Open XML（缩写：Open XML、OpenXML或OOXML），为由Microsoft开发的一种以XML为基础并以ZIP格式压缩的电子文件规范，支持文件、表格、备忘录、幻灯片等文件格式。这种文件格式基于XML（可扩展标记语言），使得文档能够以一种结构化的方式存储，并且易于解析和处理。Office Open XML规范包括几个主要部分：

1. WordProcessingML（Word文档）：用于表示Microsoft Word文档的XML格式。它包括文本、样式、表格、图像等元素的定义，使得文档的内容和格式能够被准确地表示和处理。本部分内容是系统研究的重点，将会在第4章具体展开分析。
2. SpreadsheetML（Excel表格）：用于表示Microsoft Excel电子表格的XML格式。它包括单元格、公式、图表等元素的定义，使得电子表格可以以结构化的方式进行存储和处理。
3. PresentationML（PowerPoint演示文稿）：用于表示Microsoft PowerPoint演示文稿的XML格式。它包括幻灯片、文本框、图像等元素的定义，使得演示文稿的内容和格式能够被准确地表示和处理。
4. DrawingML（绘图）：用于表示绘图元素（如图形、图表等）的XML格式。它使得绘图可以与文档、电子表格和演示文稿集成，并以可扩展的方式进行存储和处理。

OOXML规范的优点包括可扩展性、可读性和可编辑性。由于它基于XML，因此可以轻松地与其他系统集成，并且可以使用标准的XML处理工具进行处理。OOXML在2006年12月成为了ECMA规范的一部分，编号为ECMA-376；并于2008年4月通过国际标准化组织的表决，在两个月后公布为ISO／IEC 29500国际标准。但是，Word底层并未完全依据ECMA-376实现，仍然采用了许多采用了许多微软的独有规格，实际仍需结合Word的底层文档内容，参考规范文件进行对比分析得出Word文件底层的XML文件的组织方式、具体标签含义和标签数值的变化方式。

## 开发技术

### DOM4J

DOM4J是一个由Java编写的用于处理 XML 文件的开源库。DOM4J的功能十分强大，它可以将XML文件加载到内存中，并提供一系列方法来遍历、查询和操作XML文档的元素、属性和内容，控制XML的创建和修改。在性能上，DOM4J有着优异的XML解析性能，在处理大型XML文件时依然能保持高效。此外，DOM4J还支持XPath查询，这使得开发人员可以轻松地定位和访问XML文档中的特定元素和属性。

因此，本系统会使用DOM4J技术作为操作XML文件的基本库。

### Spring Boot

Spring是一个轻量级Java开发框架，最早有Rod Johnson创建，目的是为了解决企业级应用开发的业务逻辑层和其他各层的耦合问题。它是一个分层的JavaSE/JavaEE full-stack一站式轻量级开源框架，为开发Java应用程序提供全面的基础架构支持。有了Spring负责基础架构，因此Java开发者可以专注于应用程序的开发。

Spring Boot是用于创建微服务的基于Java的开源框架。它是由Pivotal Team开发的，用于构建独立的和生产就绪的弹簧应用程序。Spring Boot是对Spring的继承和发展，专注于解决Spring在配置和依赖管理上的很繁琐的问题，它帮我们预组装了Spring的一系列组件，以便以尽可能少的代码和配置来开发基于Spring的Java应用程序。Spring Boot还提供了应用监控功能，并且无需代码生成和XML配置。这使得Spring Boot不仅具有Spring一切优秀特性，而且在功能丰富性、性能稳定性和使用便捷性方面都有所提升。

总的来说，Spring Boot框架以其简化配置、内嵌容器、快速开发和微服务支持等特点，成为了现代Java应用开发的重要工具之一。因而，本系统也将采用Spring Boot作为后端开发框架。

### MySql与MybatisPlus

MySQL是一种广泛使用的关系型数据库管理系统，由瑞典MySQL AB公司开发，后被Sun Microsystems收购，最终被甲骨文公司（Oracle）收购。MySQL使用结构化查询语言（SQL）进行数据库管理，支持大量的并发连接和大量的数据存储，具有高效、稳定、易用等特点。MySQL在Web应用、数据仓库、嵌入式系统等多个领域都有广泛的应用。

MyBatis-Plus（简称MP）是一个MyBatis的增强工具，在MyBatis的基础上只做增强不做改变，简化开发、提高效率。MyBatis-Plus提供了许多实用的功能，如自动分页、全局异常处理、通用枚举注入、通用代码生成器、代码注入器等。MyBatis-Plus的核心理念是简化开发，提升效率，让开发者更加专注于业务逻辑的实现，而不用花费过多的精力在数据库操作和数据转换上。

在本系统中，我们将使用MySQL作为数据库管理系统，存储和管理用户数据、文件数据、系统配置信息等。同时，我们将使用MyBatis-Plus作为ORM框架，实现数据库操作和数据转换。

### Vue3与Vite

Vue 是一个开源的前端JavaScript框架，用于在Web上生成用户界面和单页应用程序。Vue由 Evan You 创建，于2014年发布并由Evan及其核心团队维护，专注于声明式呈现和组件组合，仅为视图层提供核心库。

Vue3是Vue.js框架的最新版本，于2020年9月正式发布。它在Vue2的基础上进行了重大改进和增强，提供了更好的性能、开发体验和更多的功能。Vue3在性能上有显著提升，比Vue2.x快1.2～2倍。它引入了Composition API，这是一种基于函数的API，让组件代码更加简洁和可复用，使开发者在编写组件时更加灵活和自由。此外，Vue3还暴露了自定义渲染API，新增了三个组件（Fragment、Teleport、Suspense），并且改进了虚拟DOM的实现和运行时编译等。Vue3更加严格地对TypeScript进行支持，提供了更加完整、准确的类型检查和错误提示，进一步提升了开发体验。

Vite是一种新型前端构建工具，用于快速搭建现代化的Vue、React或者原生JavaScript项目。能够显著提升前端开发体验。它主要由两部分组成，一是开发服务器，它基于原生ES模块提供了丰富的内建功能；二是一套构建指令，它使用Rollup打包代码，并且它是预配置的，可输出用于生产环境的高度优化过的静态资源。

在本系统中我们将采用Vue3作为前端框架，结合Vite作为构建工具，以提供高效、灵活和可维护的用户界面。

### ElementPlus

Element Plus是一款基于Vue 3.0的桌面端组件库，是Element UI的升级版本。Element Plus提供了丰富、易用的组件，如按钮、输入框、表格、对话框等，以及一套完整的主题定制方案，帮助开发者快速构建出美观、稳定的前端界面。

Element Plus继承了Element UI的优秀设计和开发理念，同时针对Vue 3.0进行了全面优化和重构。它采用了TypeScript进行开发，提供了更加严格的类型检查和更好的开发体验。此外，Element Plus还提供了丰富的主题定制选项，包括主题色、边框、字体等，让开发者能够轻松打造出符合自己项目风格的前端界面。

在本系统中，我们将使用Element Plus作为前端UI框架，提供丰富的组件和主题定制功能，让研究能更专注于业务逻辑的实现，而不用花费过多的精力在界面设计和开发上。

# 需求分析

## 系统需求概述

结合高校学位论文检测的实际需求，本系统需要实现四个方面需求：平台交互需求、格式规范检测需求、内容逻辑检测需求、生成检测报告与自动校正。

平台台交互需求是指系统需要针对不同用户，即为老师、学生、管理员提供不同的网页端和功能，让每种用户都能与系统进行交互，完成他们所需的功能。同时，提供老师与学生之间的交互功能，如学生能绑定指导老师，将论文提交给指导老师，老师也能在老师端上查看学生提交的论文。

硬性格式规范检测是指系统需要对提交的学位论文进行格式规范检测，包括论文的排版、字体、字号、页边距、页眉页脚、编号等方面的检测。系统需要基于一定的规范标准，对论文的格式进行检测。具体的格式规范也能够由用户创建、修改和删除，使得用户能用个性化的标准检测论文。此外，还需要保证格式兼容性，使得系统兼顾DOC和DOCX文件格式，并提供输出PDF文档格式的功能。

内容逻辑检测是指对于学位论文撰写中的一些涉及内容逻辑的格式规范要求进行检测。主要包括对论文参考文献引用、摘要建议字数、关键词个数、相关章节建议的篇幅、图片大小等进行检测，并给出修改意见。

生成检测报告与自动校正是指系统将检测细节展示以报告形式展现给用户，并对原文中的部分内容进行自动校正。本系统采用批注的方式给出检测报告，用批注表注相应的修改建议，以便学生进行修改和完善，与另外给出文件的报告形式相比，更方便学生对照原文修改，让用户能直接在文档中看到修改建议，提高修改效率。同时，自动校正可以让用户不必逐个修改每个格式错误，简单的格式可以交由系统自动修改，为用户节省调整格式的时间。

## 平台交互需求

根据对高校学位论文撰写流程的研究，本系统将用户分为三种：普通用户（学生）、指导老师、管理员。三种用户所需功能描述如下图所示。

表 3-1 系统用户描述

|  |  |
| --- | --- |
| **用户类型** | **功能需求** |
| 普通用户（学生） | 用户登录功能、查看论文模版规范、修改个人信息、绑定指导老师、上传论文、论文格式检测、提交论文给导师、下载检测报告、查看检测历史。 |
| 指导老师用户 | 管理论文检测模版（可创建、修改、下线、删除论文检测模版）、查看学生提交的论文检测结果、下载学生检测报告或论文原文。同时具备普通用户的登录、查看规范与论文检测功能。 |
| 管理员用户 | 管理系统账号、能批量导入账号和生成学生-老师指导关系。 |

对于普通用户和指导老师，他们既有公有的功能，如账号管理、论文检测、查看模版信息和查看检测历史，也有各自独有的功能，例如普通用户绑定指导老师并将检测论文发送给指导老师，指导老师用户可以接收学生的论文并下载其论文和论文检测报告。对于管理员，系统提供导入execl表格批量生成账号，并自动生成指导关系的功能。

### 基础功能

系统的基础功能即普通用户和指导老师用户的公有功能，如图 3-1所示，具体包括个人账号管理、论文检测、查看论文模版和查看检测历史等功能。

1. 账号管理功能

允许用户注册账号、登录账号和修改个人信息。除了基础的账号密码外，学生的账号信息包括：学生姓名、学号、年级、专业、学院、指导老师。指导老师的账号信息包括：老师姓名、教师工号、教师单位、个人描述信息。

1. 论文检测功能

论文检测功能是本系统的核心功能。用户可以将论文上传到系统，填入论文的中文标题和英文标题，选择对应的论文检测模版，系统将会根据填入的信息和选择的论文检测模版进行格式规范检测和内容逻辑检测，并生成检测结果和报告。

1. 查看论文模版

用户可以查看系统中上线的论文规范模版，根据模版的描述信息确定该模版用于哪种学位论文，能深入看到具体的模版格式要求，并且能预览每个论文模版的具体格式。

1. 查看检测历史

学生和指导老师可以在系统中查看自己的检测历史记录。检测历史记录将详细列出每次论文检测的时间、检测结果以及相应的检测报告。通过查看检测历史记录，可以下载当时检测的检测报告。



图 3-1 系统基础功能用例图

### 论文提交与论文模版管理功能

基础功能是学生与指导老师都有的与系统交互的功能，论文提交与论文模版管理则是学生用户与指导老师用户之间的交互功能。如图 3-2所示，主要包括学生绑定指导关系、向指导老师提交论文、指导老师查看收到的论文、指导老师管理论文检测模版四个功能。

1. 绑定指导关系

学生用户可以选择和特定老师组成指导关系，表示该学生用户绑定了某个老师账号作为指导老师账号，绑定后可以向指导老师提交论文。

1. 向指导老师提交论文

学生用户在设定了指导老师后，可以在进行论文检测时勾选发送到导师，则当论文检测通过后，会将论文自动发送给对应的指导老师账号，指导老师可以登录账号查看到学生发的论文。

1. 老师查看收到的论文

指导老师可以在学生论文页面查看到所有学生发来的论文，可以根据学生姓名、论文名称、检测时间等进行搜索，

1. 老师管理论文检测模板

老师可添加、删除和修改论文模板，设置模版的名称和描述信息，保存模板后，可以选择发布某些模板，发布后的模板才可以被学生使用。

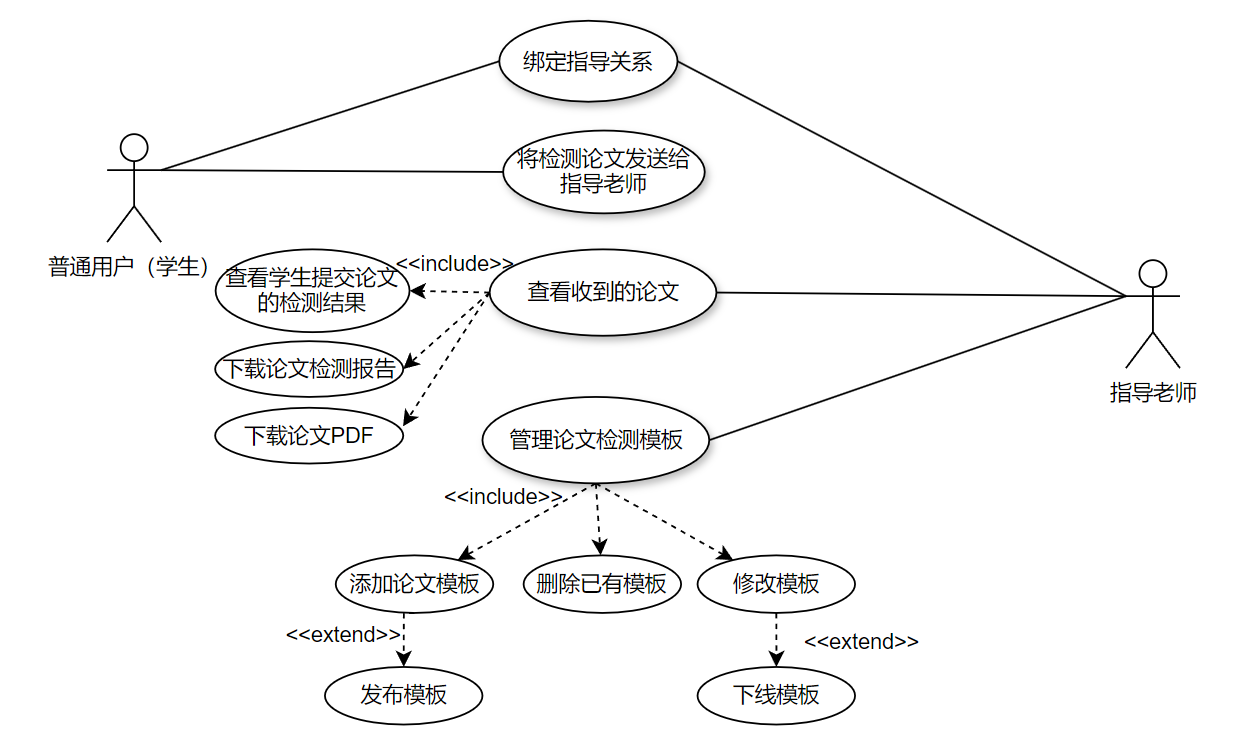


图 3-2 论文提交与论文模版管理功能

### 批量生成账号与指导关系

管理员可以根据学院提供的excel表格自动生成学生、老师账号，同时可以自动生成对应的老师-学生指导关系，可以快速创建整个院系的学生、老师账号，无需学生和老师自行注册账号。

## 内容无关格式检测

内容无关格式检测是整个系统的核心功能，系统会根据论文检测模版检测论文的格式规范。论文检测模版中需要准确定义各个论文部分的格式，对不同论文部分设定不同的格式要求，按照如图 3-3所示学位论文结构，系统的论文模版至少要定义9个部分的格式要求。同时，模版还需要补充页面设置（页边距等）、图表设置等格式要求。

对各个部分的格式要求进行拆解，具体包括标题格式检测、段落格式检测和其他相关格式。格式规范的检测不会检测具体的内容逻辑，即不会考虑具体内容、位置顺序等限制。

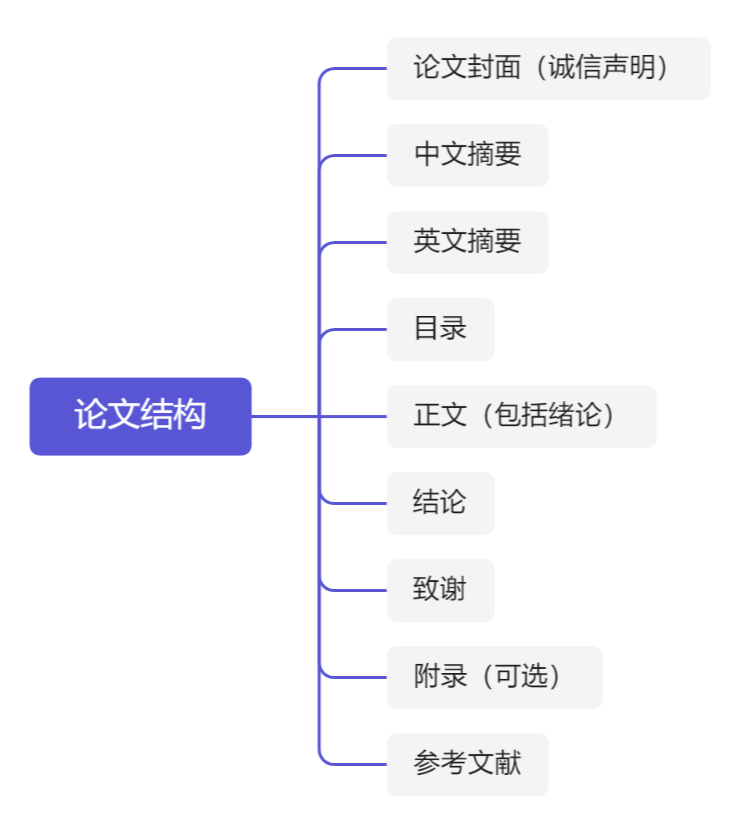


图 3-3 一般的学位论文结构

### 标题格式检测

标题格式检测主要针对论文中的各类标题，包括论文封面标题（如“诚信声明”四个字）、摘要里的中英文标题、正文部分的各级标题、目录部分的“目录”二字、结论部分的“结论”二字、致谢部分的“致谢”二字、附录部分的“附录”二字、参考文献部分的“参考文献”四字。具体检测指标和说明如下表所示。

表 3-2 标题检测指标

|  |  |
| --- | --- |
| **标题检测指标** | **说明** |
| 中文字体 | 标题的中文字体，例如宋体、楷体。 |
| 英文字体 | 标题的英文字体，例如Times New Roman。 |
| 字号大小 | 标题字号大小，例如小二、小三、四号。 |
| 对齐方式 | 标题的对齐方式，有靠左对齐、靠右对齐、居中对齐三种方式。 |
| 行距 | 设置标题的行距，例如单倍行距、1.25倍行距、1.5倍行距。 |
| 段前距 | 设置标题前的间隔距离，例如0.5行、1行。 |
| 段后距 | 设置标题后的间隔距离，例如0.5行、1行。 |
| 加粗 | 设置标题是否加粗。 |

### 段落格式检测

段落格式检测主要针对论文中各部分标题下的正文内容，包括诚信声明内容、摘要内容、绪论内容、正文各级标题下内容、结论内容、致谢内容、参考文献内容。具体检测指标和说明如下表所示。

表 3-3 段落检测指标

|  |  |
| --- | --- |
| **段落检测指标** | **说明** |
| 中文字体 | 段落的中文字体，例如宋体、楷体。 |
| 英文字体 | 段落的英文字体，例如Times New Roman。 |
| 字号大小 | 段落字号大小，例如四号、小四。 |
| 段落缩进 | 段落内容的缩进，控制段落缩进的长度。如每段缩进两个字符。 |
| 段落行距 | 设置段落的行距，例如单倍行距、1.25倍行距、1.5倍行距。 |
| 加粗 | 设置段落是否加粗。 |

### 其他格式检测

除了标题和段落格式这种论文每个部分都会涉及到格式要求外，一些部分特有的格式也要考虑，例如摘要的格式、参考文献的格式，以及一些页面设置、编号设置都是需要系统进行检测的。具体如下。

1. 摘要格式

中英文摘要都应该满足先有标题，然后摘要正文，最后是关键词，顺序不能错。同时要对摘要的前导词进行检测，例如中文摘要的摘要正文前要有前导词“[摘 要]”，关键词前要有前导词“[关键词]”，系统也要对这种前导词的格式进行检测。每个关键词中间应该用“；”隔开，对于错误的分割词，系统也要等够检测出来。英文摘要同理，也应该有前导词“Abstract：”和“Keywords”，以及分割字符“；”。

1. 参考文献引用格式

在正文中引用参考文献，必须使用中括号框选引用论文编号，并以上角标的格式显示，如“[1]”。引用多个编号时，多个编号之间用“，”隔开，同时引用多个连续的参考文献时，可以使用“~”符。

1. 页面尺寸格式

系统需要检测页面尺寸。在页面尺寸上，默认A4纸张大小，同时需要设置页边距，包括页上边距、页下边距、页左边距、页下边距。

1. 页眉页脚格式

系统需要检测页眉页脚的格式。在论文模版中需要规定页眉格式，包括页眉的字体类别、字号大小、对齐方式。对于页脚，需要插入页码，规定页码数字的位置、字体类别、字号大小以及是否区分奇偶页。

1. 图注表注格式

系统需要检测图注表注的格式。对图注表注的字体类别、字号大小、对齐方式等格式进行检测。

1. 编号格式

系统需要检测论文的编号格式。针对论文中正文（包括绪论）部分的内容，论文检测模版需要规定一级、二级、三级标题编号的样式，如果原文的标题编号样式不对，需要进行自动更正或给出提醒。同时，对于参考文献，需要检测每条参考文献是否都以正确的编号格式，即采用中括号包裹的形式进行编号。

## 内容相关格式检测

内容相关格式检测与3.3节的内容无关格式检测不同，这部分的格式检测需要涉及对内容的检测，例如论文缺少某部分内容，或某部分内容有冗余，以及对内容篇幅、文献引用方面的要求。这里有些要求并非是硬性的，即可以满足也可以不满足，但是建议满足，这些内容逻辑检测对保证整篇论文的规范性同样十分重要。主要包括以下内容。

1. 内容缺失或冗余检测

系统会根据图 3-1展示的一般论文结构来检查整篇论文的结构完整性。如果论文中有重复的部分，例如有两个中文摘要，或有某些部分缺失了，例如缺少了关键词，系统都要在检测报告中指出。

1. 内容篇幅限制

多数学位论文的撰写要求中都会涉及篇幅，即使没有明确指出，指导老师在帮学生改论文时，也会要求学生满足一定的内容篇幅限制。这些内容篇幅也需要系统进行记录和检测。下表列举了常见的内容篇幅限制。

表 3-4 论文常见的篇幅限制

|  |  |
| --- | --- |
| **篇幅限制** | **举例说明** |
| 摘要篇幅的限制 | 例如摘要内容不要超过两页。 |
| 结论篇幅的限制 | 例如结论内容不要超过两页。 |
| 致谢篇幅的限制 | 例如致谢内容不要超过一页。 |
| 参考文献条数的限制 | 例如参考文献条数应当不少于8条。 |
| 关键词个数的限制 | 例如关键词个数不多于6个，不少于2个。 |

1. 文献引用限制

列出的每条参考文献，需根据一定的规范组织，例如每条参考文献应按照国标GB/T 7714-2015标准列出。同时系统需要检测是否有参考文献全篇都没有被引用，以及是否正文中引用了无效的参考文献，即检测未引用和无效引用。此外，对参考文献部分每条参考文献的顺序也要按照在正文中的出现顺序进行组织排序。

1. 图注表注限制

除了图注表注的格式，系统还需要检测是否有图片缺失了图注，是否有表格缺失了表注，是否有无效的图注表注（即没有对应图片或表格的注释）。同时，系统还需要确保图注表注出现在了正确的位置，对于图注，需要在图片的正下方；对于表注，需要在表格的正上方。

1. 图片大小限制

系统需要检测论文中插入的图片的大小，以确保图片尺寸不至于过大或过下从而影响阅读。当出现过小或过大的图片时，需要为用户给出提示信息。

1. 页眉页脚的内容限制

系统需要检测页眉页脚的内容是否正确。例如，对于页眉，一般学位论文会要求在页眉上展示本篇论文的中文标题，若检测发现页眉内容不正确，系统要给出提醒或自动更正；对页脚，主要是需要检测开始展示页码的页面位置是否符合规范，同时保证页脚除了页码外无其他无效字符。

1. 目录内容缺失

系统还需要检测目录的内容是否完整。对于正文中各章节的一级、二级、三级标题，系统需要确保目录中的内容包含这些标题，以及标题与实际论文中的标题一致。如果目录中的某些标题缺失系统需要给出提醒，以便作者及时更正。

## 生成检测报告与自动校正

系统在完成检测后，需要给出检测结果，并以批注的形式给出检测报告。系统需将文件现有的批注删除，然后根据系统格式检测的结果，在特定文段上添加批注，让用户可以根据批注导向到相关文段内容，或根据文段中被高亮的内容导向相关批注，让论文原文与检测报告能在同一个Word文件中，以此方便用户自己根据批注对论文内容进行检查和修改。

自动校正是指系统能自动修改格式的地方系统自动修改，用户仅需确认检验即可。自动校正的内容需要包括3.3节中的硬性格式要求和部分3.4节提到的与内容逻辑相关的格式。

为了有便于用户阅读的检测报告，并结合自动校正功能，系统的检测报告需要有4种类型的批注。

1. 概述批注

位于文件开头，给出整篇检测报告的概述。给出全文共有多少批注建议，给出每种批注类别的数量。

1. 自动修改批注（对应可自动修改的硬性格式要求）

系统检测到硬性格式要求，且该格式要求可以自动修改时，系统会自动进行校正，并在校正处留下自动修改批注。对此类批注，用户仅需简单确认一下即可。

1. 待修改批注（对应不可自动修改的格式要求）

系统检测到必须要修改的格式要求，且该格式无法自动修改，系统会在此处留下待修改批注，提醒用户进行修改。

1. 建议修改批（对应软性格式要求）

系统检测到论文不符合论文检测预设的格式要求，但该格式并不是强制性的，则会留下建议修改批注，建议用户进行修改。

# WordProcessingML文档分析

WordProcessingML是OOXML的重要部分，是用于描述Word文档（DOCX格式）的XML格式，其内部规定了DOCX底层的文件组织方式和内容组织方式。它包括文本、样式、表格、图像、批注等各种DOCX文档元素的定义。理解和分析DOCX文件底层内容的组织方式，是进行格式检测的基础。

本章将从DOCX的文件组织结构展开，分析DOCX底层如何控制文档内容的格式。因整个格式规定十分庞大，以下只列出研究分析得出的与论文格式检测相关的重要内容。

## 文件组织结构

DOCX文件本质是一个压缩包文件，将后缀改为“.zip”后即可解压文件。使用解压缩软件解压DOCX文件后，可以看到DOCX文件的底层文件组织结构。

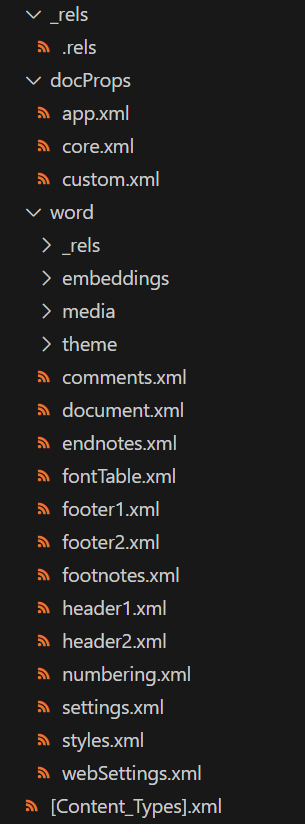


图 4-1 DOCX文件解压后的文件结构

DOCX文件解压后主要有xml和rels文件。xml主要存储文档内容和定义样式，rels文件作为元数据文档，存储文档的关联信息。rels文件还负责告诉Word文档编辑器在何处查找文档内容和其他参考的信息，内部每个关系都由唯一的关系 ID 标识，并将引用的 XML 文件指定为目标。上图4-1中各个文件夹或文件的作用描述如下表所示。

表 4-1 DOCX内部各文件的作用描述

|  |  |  |
| --- | --- | --- |
| **文件名称** | **类型** | **作用** |
| \_rels | folder | 存储文档的关联信息。 |
| docProps | folder | 存储文件属性信息，包括app.xml、core.xml、custom.xml文件。 |
| docProps/app.xml | file | 描述文档的其他属性，文档类型，版本，只读信息，共享，安全属性等特定的文件属性。 |
| docProps/core.xml | file | 描述文件的创建时间、标题、主题和作者等设定的xml约定文档格式的通用文件属性。 |
| docProps/custom.xml | file | 描述文档的个性化属性。 |
| word | folder | 存储文档内容，是整个DOCX文档的核心。 |
| word/\_rels | folder | 存储文档内容的关联信息。 |
| word/embeddings | folder | 存储文档的嵌入式文件。这些嵌入的对象可以包括其他Office文档、图像、音频、视频等。 |
| word/media | folder | 存储文档的媒体文件，主要是图片 |
| word/theme | folder | 存储文档的主题信息，主要是Office Word中预设的主题。 |
| word/document.xml | file | 存储文档的正文内容。文档中能看到的所有内容，在这里都有直接或者间接的记录。 |
| word/comments.xml | file | 存储文档的批注。 |
| word/endnotes.xml | file | 存储文档的尾注。 |
| word/fontTable.xml | file | 存储文档的字体表。 |
| word/footer.xml | file | 存储文档页尾内容。 |
| word/header.xml | file | 存储文档页眉内容。 |
| word/numbering.xml | file | 存储和控制文档的编号。 |
| word/settings.xml | file | 存储文档内容的设置。 |
| word/styles.xml | file | 存储和定义文档的样式集。 |
| word/webSettings.xml | file | 用于指定在将网页另存为 Word 文档或将 Word文档另存为网页时适用的文档级属性的设置 |
| [Content\_Types].xml | file | 描述文档中包含的各种媒体类型和其对应的扩展名，以及与这些媒体类型相关的默认处理方式。 |

## 控制基本样式

### 基本内容和样式的组织结构

DOCX文档的基本样式控制主要集中在word/document.xml文件和word/styles.xml文件中。在document.xml文件中，会存储文档的具体内容，导入在styles.xml中预定义的样式，针对一般的段落和标题，基本内容和样式的组织结构如下。



图 4-2 document.xml内容和样式的基本组织结构

如图所示，<w:document>标签是document.xml的根标签，<w:body>标签表示存储文档的正文内容，<w:p>表示存储内容是一个文段（包括标题、段落和插入内容），<w:pPr>描述<w:p>表示文段的属性，<w:r>表示存储文段中的一个文块（即文段的一部分内容），<w:rPr>描述该文块的属性。<w:t>则具体存储文块的文本内容。

### 分层级的样式控制

DOCX对文档内容的样式控制并不集中在单独一个标签，而是具有多层级的样式控制，包括预定义样式、文段样式、文块样式三层，各层样式之间的样式控制有重合，样式优先级按顺序由低到高，对应的控制范围也逐渐变小。当各层样式有冲突时，以优先级高的样式为准。根据图4-2描述的结构，可以看到，优先级越高的样式在document.xml的越深层位置描述。各层样式控制分析如下。

1. 预定义样式

预定义样式是在styles.xml文件中定义的，日常使用Office Word在“开始”功能栏，右上角展示的样式就是预定义样式。

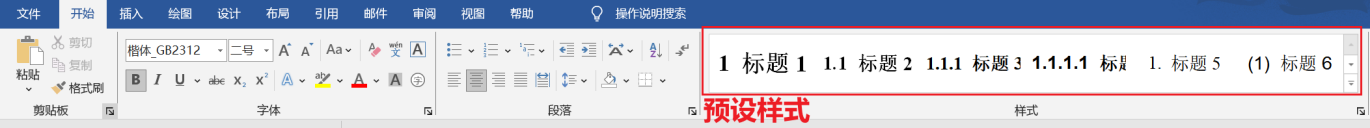


图 4-3 预定义样式

下图是在styles.xml文件中定义样式的一个示例，这段内容定义了一个名为“Normal”的预定义样式，并指定了样式的styleId值为a。在document.xml文件中导入该预定义样式时需要根据styleId值来指定该样式。此外，预定义样式可以被继承，例如在定义其他预定义样式时，可以继承另一个已有的样式，当已有样式与当前样式有控制上重合时，以当前定义样式为准。

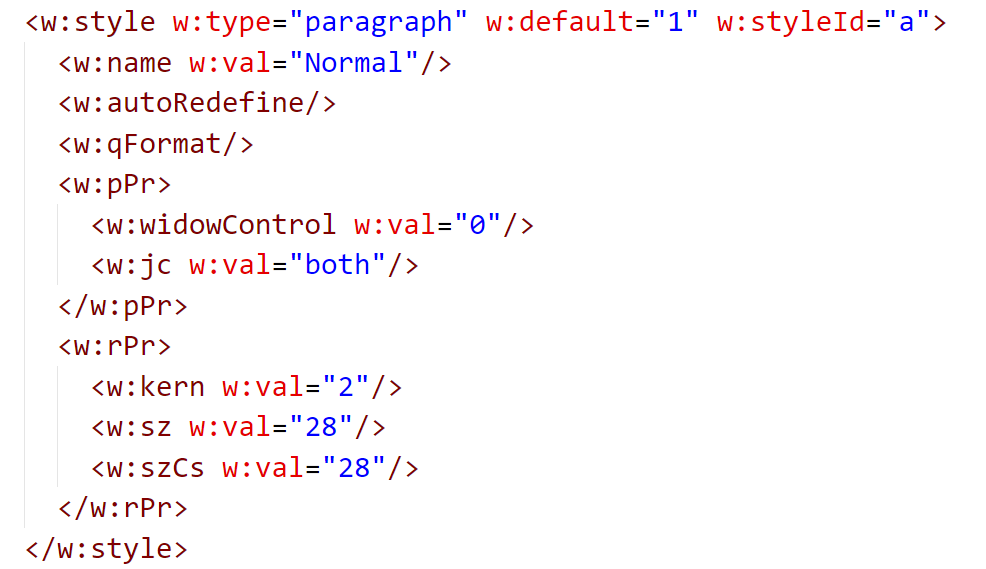


图 4-4 一个预定义样式的示例

1. 文段样式

定义在<w:pPr>标签中，表示控制本文段的样式，控制范围仅局限在一个文段内，即<w:p>内。

1. 文块样式

定义在<w:rPr>标签中，表示控制文块的样式，控制范围仅局限在一个文块内，即<w:r>内。

### 常见样式控制标签

样式控制标签是控制样式的最小单位，只控制某个单独的样式，例如加粗、字体大小、对齐方式等。下面列出格式检测需要用到的常见的控制标签。

表 4-2 常见样式控制标签

|  |  |
| --- | --- |
| **标签名** | **作用** |
| w:jc | 控制对齐方式，包括左对齐、右对齐、居中对齐和两端对齐。 |
| w:spacing | 控制段落间距，包括段前距、段后距、段间距，并指定间距距离。 |
| w:rFonts | 控制文本字体类别。 |
| w:sz | 控制文本大小 |
| w:b | 控制内容是否加粗 |
| w:ind | 控制内容缩进。 |
| w:bidi | 控制文本的显示方向，可控制从左到右书写和从右到左书写。 |
| w:kern | 控制字符间距。 |
| w:lvl | 控制编号等级（控制序号大小）。 |
| w:numId | 控制编号样式 |
| w:vertAlign | 控制角标样式，包括上角标、下角标。 |

值得注意的是，在一个样式标签中，可以设置多种属性，且可能会同时存在多个属性值，而这些属性值之间也具有优先级（或覆盖）关系。例如，在控制段落间距的标签spacing中，属性w:before和w:beforeLines都是用于控制段后距，而当存在这两种属性时w:before属性失效（被覆盖）。因而在使用系统调整样式时，要先寻找到优先级最高的样式。

## 控制页面尺寸与分节

DOCX的页面尺寸与分节都在document.xml文件<w:sectPr>标签中控制。页面尺寸能保证页面内容在合理的范围内展示，分节是将所有页面划分为几个部分，每个部分单独控制页眉页脚和页面尺寸。

WordProcessingML中规定，<w:body>标签的最后一个子元素必须是<w:sectPr>标签，该标签控控制默认节的设置，如果有设置多个节，那么该标签就控制最后一节的设置。如果存在多个节，那么就会有多个<w:sectPr>标签，除了最后一个标签外，其它标签都在<w:p>标签内。每在<w:p>标签中添加一个<w:sectPr>标签，即相当于在页面上插入了一个“分节符”。下图是一个<w:sectPr>标签的示例。

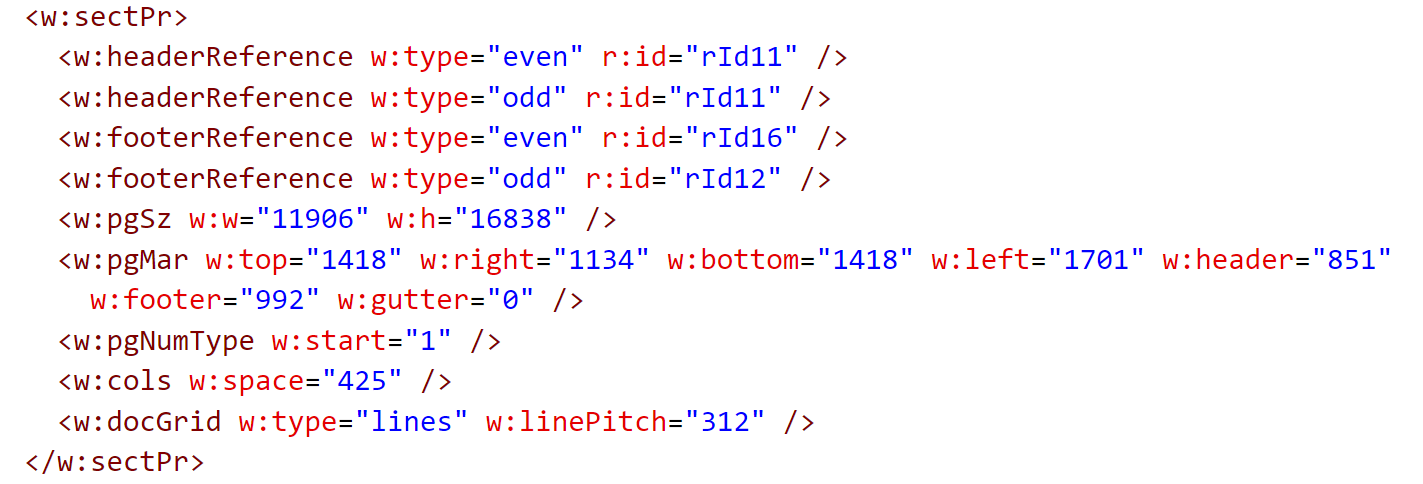


图 4-5 <w:sectPr>标签示例

<w:sectPr>标签中控制长度的属性值满足“567 = 1cm”，在上图示例中，<w:pgSz>规定了本节的页面大小，即21cm×29.7cm（标准A4纸张大小）。<w:pgMar>规定了页边距，包括上边距、下边距、左边距、右边距、页眉边距、页脚边距。

## 控制页眉页脚

页眉页脚的内容在header.xml和footer.xml文件中定义，文本内容分别定义在标签<w:hdr>和标签<w:ftr>（包括页码的定义）中。文件定义好后需要在rels文件和 [Content\_Types].xml文件中定义关系和类型，随后可以在document.xml文件中通过<w:sectPr>标签引用定义好的页眉页脚。

有些论文格式会要求控制奇偶页不同。DOCX中控制奇偶页首先得在settings.xml文件添加<evenAndOddHeaders>标签，表示本文档区分奇偶页，随后可以在引用页眉页尾文件时附带上属性值w:type，来控制本节奇偶页分别引用的页眉页尾。

如图 4-5 所示的示例，标签<w:headerReference>引用了关系id（在rels文件定义）为rId11的页眉，w:type="even"表示设置在偶数页。同理，<w:footerReference>引用了关系id为rId12的页尾，w:type="odd"表示设置在奇数页。

## 控制编号

WordProcessingML规定，编号项的样式定义在numbering.xml文件。<w:numbering>为numbering.xml的根标签，在其子标签<w:abstractNum>中定义编号样式，例如“1, 2, 3, …”，“a, b, c, …”，“I, II, III, …”等，<w:lvl>标签可以分层级定义每一级编号的样式，例如“一、1、（1）”，一级标题为“一、”，二级标题为“1、”，三级标题为“（1）”，同时可以指定该编号样式下的文段要使用的默认样式，用<w:pStyle>指定。下图是一个编号样式定义示例。



定义编号样式后，需要在<w:numbering>标签内用<w:num>标签指定编号样式的numId，之后便可在document.xml中具体文段的<w:pPr>标签内使用<w:numPr>标签引用编号样式，即说明该部分文段使用特定的编号样式。例如下图所示，表示该文段被设置为一级编号（<w:ilvl>属性值为0），且使用的编号样式id为0（<w:numId>属性值为0）。

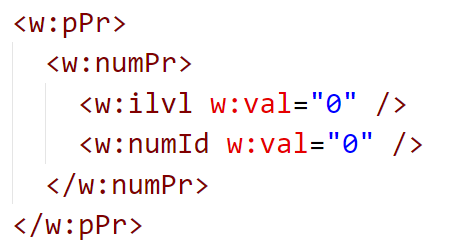


图 4-6 某文段指定编号样式的示例

## 控制图表和文本框

图片、表格和文本框这类特殊内容，在DOCX中也是存储在document.xml文件中，其中图片这类媒体文件的本体存储在word/media文件夹中，并在rels文件中定义关系id，之后便可在document.xml中引用。

1. 控制图片

WordProcessingML规定，图片仍然属于文段内容，因而控制图片插入的内容在<w:p>标签内容下。控制图片的标签有两类，一类是<w:drawing>，另外一类是<w:pict>，前者是手动添加图片时会生成的标签，后者是采用外部程序在DOCX文件中生成图片时会产生的标签，或是在将DOC文件转化为DOCX文件时产生。在实际的控制图片标签中，为了方便Word编辑软件渲染图片，会附带很多信息，这些信息很多是提供给Word编辑软件的信息，并不实际控制图片样式。在把一些次要信息剔除后，控制图片标签<w:drawing>的结构大致如下。



图 4-7 控制图片插入

在上面的实例结构中，<wp:inline>指示图片以“内联”方式插入，这意味着图片将直接嵌入到文本流中，而不是浮动在文本周围。这个元素还包含几个属性distT="0"、distB="0"、distL="114300"、distR="114300"，这些属性定义了图片与周围文本的距离（上、下、左、右），单位为磅（1/20英寸）。<a:graphic>：这是一个通用的图形容器，用于包含图形描述信息。xmlns:a和xmlns:pic定义了所使用的命名空间。<pic:pic>是图片的容器元素标签，它定义了图片的具体内容和属性。<pic:blipFill>中的<a:blip r:embed="rId7" />指定了图片的实际数据源。r:embed="rId7"是一个引用关系ID，指向实际的图片文件（存储在media文件中）。<a:xfrm>描述了图片的位置和大小。<a:off >定义了图形在页面上的偏移量（x，y坐标），这里都设为0。<a:ext>定义了图形的大小，cx 和 cy 分别表示宽度和高度，单位是EMU（英制测量单位，每英寸914400 EMU），大概是1厘米 = 360000EMU。

1. 控制表格

WordProcessingML规定，表格是与文段同等级的内容，使用<w:tbl>表格标签，内部可以嵌套<w:tbl>标签或<w:p>标签。<w:tbl>定义表格的内容结构和说明如下。

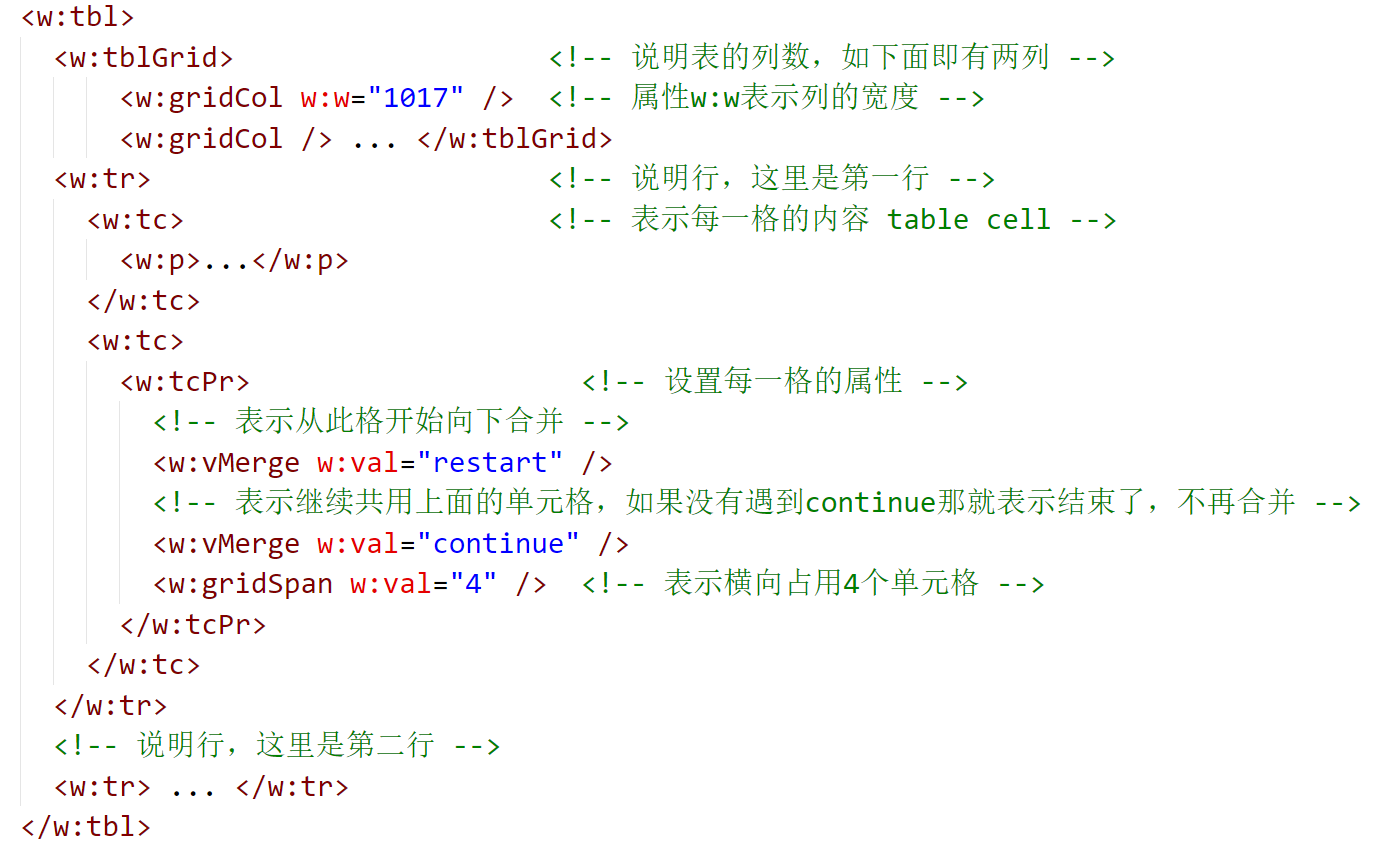


图 4-8 表格的定义与说明

1. 文本框

文本框在通常的学位论文撰写中不会使用，但根据WordProcessingML规范，文本框也同样用<w:drawing>控制，容易在检测时与图片混淆。文本框控制的区别在于其<w:drawing>标签内部没有<pic:pic>标签，且通常会有<wp:docPr name="文本框"/>的描述标签，定义整个元素的名字。

## 控制批注

WordProcessingML规定，DOCX文档批注内容定义在comments.xml文件中。要使得comments.xml的内容生效，需要首先在rels文件中定义关系id，并在[Content\_Types].xml中定义类型，最后在document.xml中引用批注。

1. 在comments.xml中定义批注内容

DOCX的每一条论文都存储在<w:comment>标签中，用标签属性设置批注的编号、批注人和批注时间，用文段标签<w:p>存储批注的具体内容。此外，批注内容的样式可以引用styles.xml中定义的批注样式。下图是批注结构示例。



图 4-9 批注标签的结构示例

1. 定义comments.xml文件的关系id和类别

如下图所示，在rels文件和[Content\_Types].xml文件添加对应的关系id和文件类型。

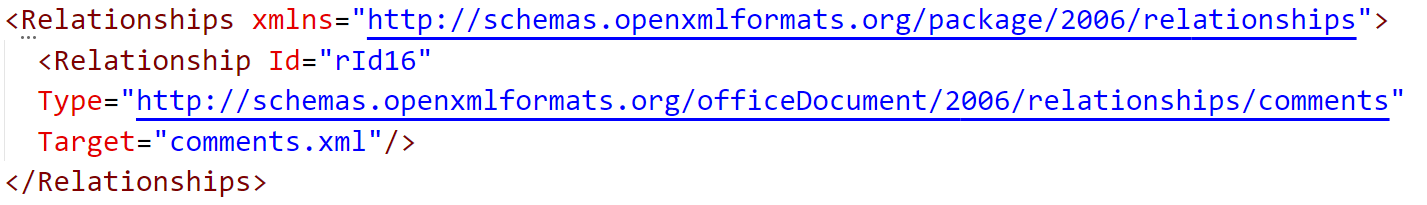


图 4-10 定义comments.xml文件的关系id

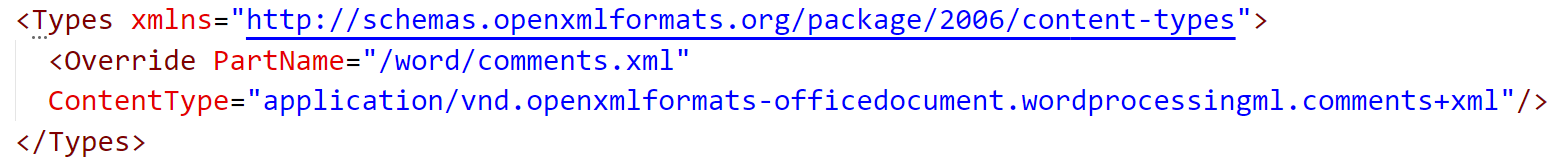


图 4-11 定义comments.xml的类别

1. 在document.xml中引用批注

定义好批注内容后，需要在document.xml文件中引用批注，设置该提条批注的位置，以及是否指定某段文字。

若只不选中内容，只是在某处插入批注，那么只需要在对应的<w:r>标签后插入标签<w:commentReference>，并用w:id属性指定批注的编号。

若需要选中内容，则需先将选择内容用<w:commentRangeStart>和<w:commentRangeEnd>标签包裹，标识选择内容的开始位置和结束位置，并在结束标签后插入<w:commentReference>标签。下图是一个选中内容的范围批注示例。

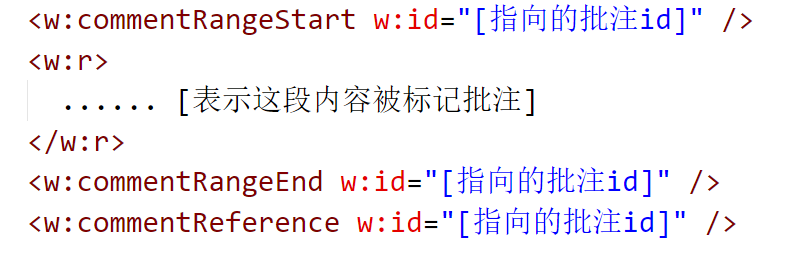


图 4-12 选中内容的批注示例

# 系统设计

本节会先介绍整个系统的总体结构和数据库设计，然后将系统的功能设计拆分成平台交互功能设计、内容无关格式检测功能设计、内容相关格式检测功能设计和报告生成与自动校正功能设计。

## 系统总体结构

结合第三章对本系统需求进行的整理和分析，第四章对DOCX底层的文件组织方式和内容组织方式的整理，本系统将会采用B/S结构构建Web应用程序。系统总体架构如下图所示。

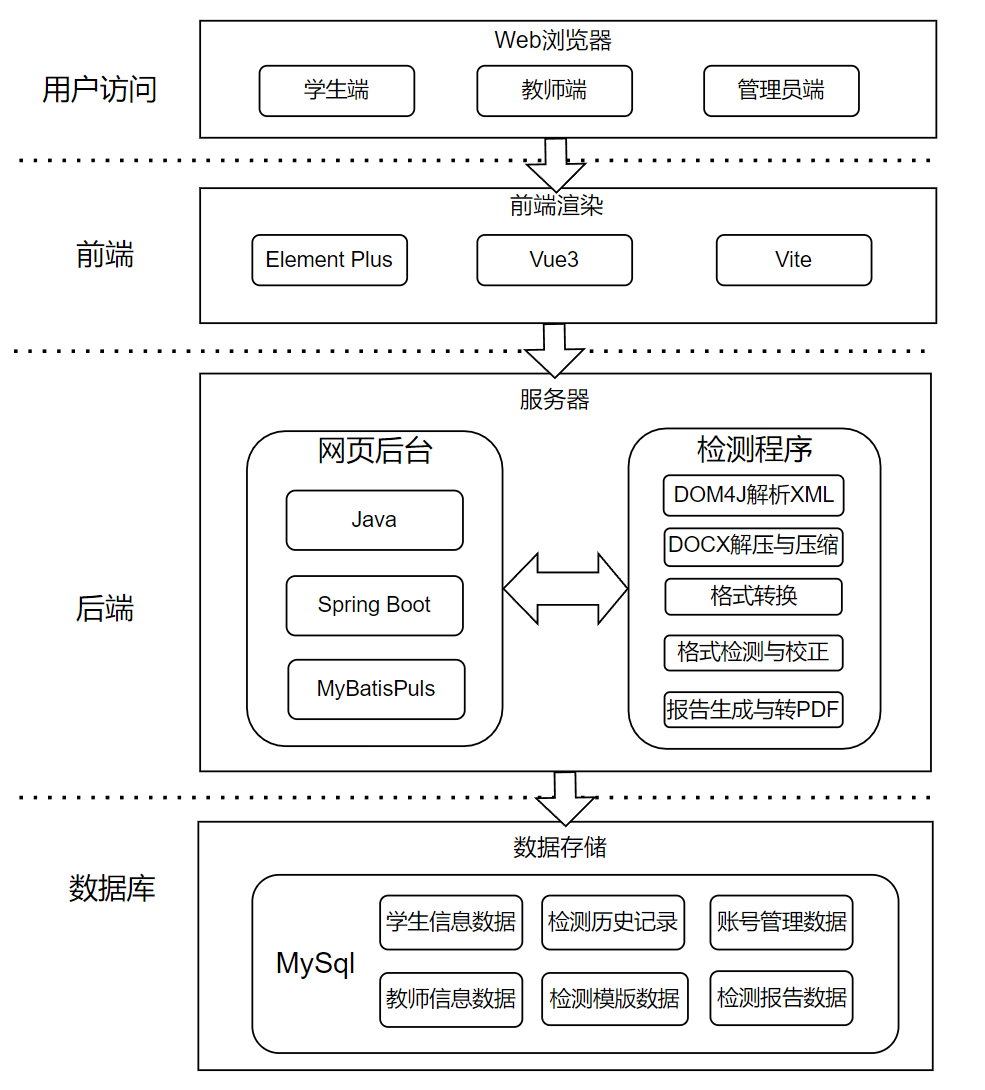


图 16 系统总体架构

## 数据库设计

数据库设计是系统设计的重要部分，它负责存储和管理系统所需的各种数据。

（TODO ）e-r图，看情况放。

对于本论文检测系统，主要设计有以下六个表。

1. 账号表

账号表用于存储账号信息，如账号名称、密码、用户类别等。账号名称唯一标识一个账号，密码字段存储的是密码的哈希值。表中的这些信息用于区分用户类别，完成用户登录的功能。

表 5-1 账号表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名 | 数据类型 | 能否为空 | 是否主键 | 是否外键 | 描述 |
| username | varchar(100) | NO | YES | NO | 用户账号名 |
| password | varchar(200) | NO | NO | NO | 账户密码的哈希值 |
| role | varchar(20) | NO | NO | NO | 账户角色：学生STUDENT、教师TEACHER、管理员ADMIN |

1. 学生信息表

学生信息表用于存储学生的基本信息，包括学生账号名、学生姓名、学生学号、年级、专业、所属学院和指导老师。每个学生有唯一的学号，学生可以没有指导老师。

表 5-2 学生信息表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名 | 数据类型 | 能否为空 | 是否主键 | 是否外键 | 描述 |
| username | varchar(100) | NO | NO | YES | 用户账号名 |
| student\_id | varchar(50) | NO | YES | NO | 存储学生学号 |
| name | varchar(50) | YES | NO | NO | 存储学生姓名 |
| grade | varchar(50) | YES | NO | NO | 存储学生年级 |
| major | varchar(50) | YES | NO | NO | 存储学生 |
| departmant | varchar(100) | YES | NO | NO | 存储所属学院 |
| Instructor | varchar(100) | YES | NO | YES | 存储指导老师账号名 |

1. 教师信息表

教师信息表用于存储教师的基本信息，包括教师账号名、教师姓名、教师工号、教师所属部门和简介信息。每个教师有唯一的教师工号。

表 5-3 教师信息表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名 | 数据类型 | 能否为空 | 是否主键 | 是否外键 | 描述 |
| username | varchar(100) | NO | NO | YES | 账号名 |
| name | varchar(50) | NO | NO | NO | 老师姓名 |
| teacher\_id | varchar(50) | NO | YES | NO | 教师工号 |
| department | varchar(50) | YES | NO | NO | 所处单位 |
| description | varchar(255) | YES | NO | NO | 信息简介 |

1. 检测记录表

检测记录表用于记录每次论文检测的结果，每次检测都有唯一的检测id，每条记录会存储进行检测的用户账号、使用的模版id、检测结果、检测论文中英文名称、检测结果报告的名称。如果是学生，还会记录本次检测是否发送给指导老师，以及该指导老师的账号名。

表 5-4 检测记录表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名 | 数据类型 | 能否为空 | 是否主键 | 是否外键 | 描述 |
| record\_id | int | NO | YES | NO | 检测编号 |
| username | varchar(100) | NO | NO | YES | 账号名 |
| template\_id | varchar(50) | NO | NO | YES | 模板id |
| detect\_time | varchar(50) | NO | NO | NO | 检测时间 |
| status | int | NO | NO | NO | 检测状态：0不通过，1通过，2通过但可修改 |
| paperName | varchar(50) | NO | NO | NO | 论文中文名 |
| paperEnglishName | varchar(50) | NO | NO | NO | 论文英文名 |
| resultFileName | varchar(50) | NO | NO | NO | 检测结果文件 |
| resultPDF | varchar(50) | YES | NO | NO | 检测结果文件pdf，只有当通过检测后才会有 |
| isSendToTeacher | int | NO | NO | NO | 1为是，0为否 |
| teacher\_username | varchar(100) | YES | NO | YES | 存储当时指导老师账号名 |

1. 模版信息表

模版信息表用于记录每个模版的基本信息，包括模版id、模版名称、模版的创建者、模版的创建时间、模版更新时间、模版的状态和模版的描述信息。模版创建者可以在教师端登录后修改这些信息。

表 5-5 模版信息表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名 | 数据类型 | 能否为空 | 是否主键 | 是否外键 | 描述 |
| template\_id | varchar(100) | NO | YES | NO | 模板编号 |
| template\_name | varchar(100) | NO | NO | YES | 模板名称 |
| creator | varchar(50) | NO | NO | NO | 创建人 |
| create\_time | varchar(50) | NO | NO | NO | 创建时间 |
| update\_time | varchar(100) | NO | NO | NO | 最近更新时间 |
| status | int | NO | NO | NO | 0表示未发布，1表示已发布 |
| description | varchar(255) | NO | NO | NO | 模板描述 |

1. 模版规范信息表

模版规范信息表详细记录了每个模版所规定的格式规范，包括论文诚信声明、目录、摘要、正文、结论、致谢、参考文献等各个部分的标题格式要求与段落格式要求，以及页面设置、图注表注设置、篇幅限制等要求。每个模版一一对应一条模版规范信息记录，模版规范信息表通过模版id与模版信息表关联。因为该表的字段数较多（共172条字段），下面仅列出几个代表性部分的字段，同时值保留字段名和描述信息，省略数据类型、是否为空等次要信息。

表 5-6 模版规范信息表（节选）

|  |  |  |
| --- | --- | --- |
| 控制范围 | 字段名 | 描述 |
| 模版id | template\_id | 绑定的模板id |
| 诚信声明 | soh\_flag | “诚信声明”标志位，0无，1有 |
| soh\_seq | 次序，默认为1 |
| soh\_content | 诚信说明内容 |
| soh\_h\_font\_type | "诚信声明"标题字体类别 |
| soh\_h\_font\_english\_type | 西文字体类别 |
| soh\_h\_font\_sz | 字体大小 |
| soh\_h\_jc | 对齐方式left，right，center |
| soh\_h\_bold | 字体加粗，0无，1有 |
| soh\_h\_afterline | 段后间隔，100为1行 |
| soh\_h\_beforeline | 段前间隔，100为1行 |
| soh\_h\_line | 段间距，240为1行 |
| soh\_p\_font\_type | "诚信声明"段落字体类别 |
| soh\_p\_font\_type\_english | 西文字体类别 |
| soh\_p\_font\_sz | 字体大小 |
| soh\_p\_ind | 首行缩进 |
| soh\_p\_line | 段间距 |
| soh\_p\_bold | 段落正文是否加粗，0无1有 |
| 中文摘要 | aoc\_flag | 中文摘要是否有 |
| aoc\_seq | 中文摘要次序 |
| aoc\_prefixFont | 前置词字体 |
| aoc\_isPrefixBold | 前置词是否加粗 |
| aoc\_recommendedMaxContentLength | 摘要的推荐最大字数 |
| aoc\_recommendedMinContentLength | 摘要的推荐最小字数 |
| aoc\_recommendedMaxKeywordsCount | 关键词的推荐最大个数 |
| aoc\_recommendedMinKeywordsCount | 关键词推荐的最小个数 |
| aoc\_h\_font\_type | 标题字体类别 |
| aoc\_h\_font\_english\_type | 西文字体类别 |
| aoc\_h\_font\_sz | 字体大小 |
| aoc\_h\_jc | 对齐方式left，right，center |
| aoc\_h\_bold | 字体加粗，0无，1有 |
| aoc\_h\_afterline | 段后间隔，100为1行 |
| aoc\_h\_beforeline | 段前间隔，100为1行 |
| aoc\_h\_line | 段间距，240为1行 |
| aoc\_p\_font\_type | 段落字体类别 |
| aoc\_p\_font\_type\_english | 西文字体类别 |
| aoc\_p\_font\_sz | 字体大小 |
| aoc\_p\_ind | 首行缩进 |
| aoc\_p\_line | 段间距 |
| aoc\_p\_bold | 段落正文是否加粗，0无1有 |
| 页面设置 | pageSetting\_headerContent1 | 页眉内容1 |
| pageSetting\_headerContent2 | 页眉内容2 |
| pageSetting\_topMargin | 页上边距 |
| pageSetting\_bottomMargin | 页下边距 |
| pageSetting\_leftMargin | 页左边距 |
| pageSetting\_rightMargin | 页右边距 |
| pageSetting\_headerMargin | 页眉边距 |
| pageSetting\_footerMargin | 页脚边距 |
| pageSetting\_oddEvenPage | 奇偶页设置，0无，1有 |
| 图注表注 | caption\_fontName | 图表注释字体 |
| caption\_fontEnglishName | 图表注释英文字体 |
| caption\_fontSize | 图表注释字号 |
| caption\_picSzWidthMin | 图片最小宽度 |
| caption\_picSzWidthMax | 图片最大宽度 |
| caption\_picSzHeightMin | 图片最小高度 |
| caption\_picSzHeightMax | 图片最大高度 |
| caption\_recommendNum | 推荐的图表数量，这个数量是最小值，可以多于这个数量。 |

## 平台交互功能设计

系统的平台交互功能主要靠网页端的功能实现，设计平台交互功能即设计不用用户网页端的功能。根据系统需求分析，设计三个网页端：普通用户端、教师端和管理员端。将需要实现的功能按模块进行拆分，设计出如下图所示的系统网页端功能架构图。

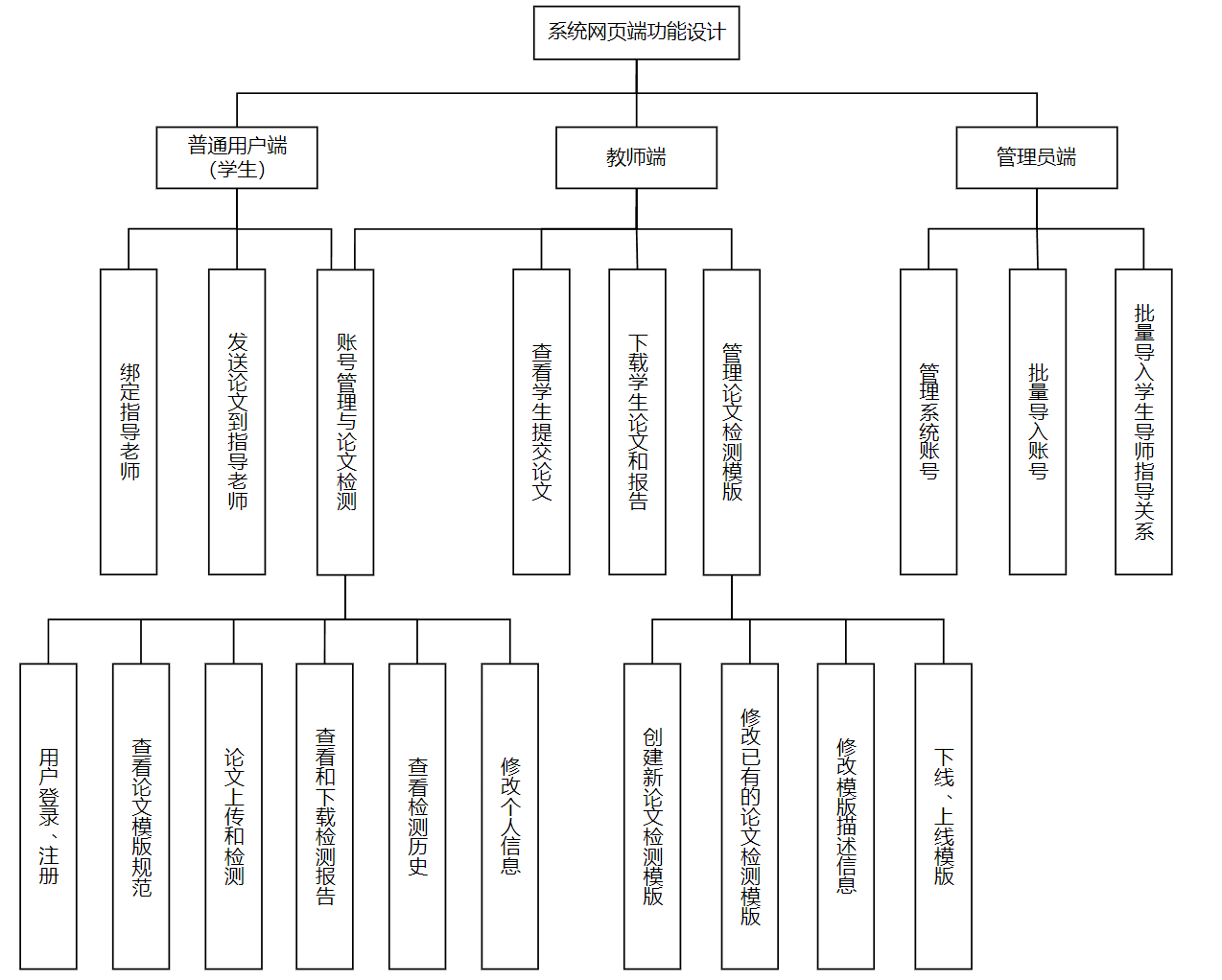


图 5-2 系统网页端功能架构图

### 普通用户端

普通用户是系统人数最多的用户，因而普通用户端承载着系统最基础和最常用的功能。主要有以下四个部分，其中账号管理与论文检测是普通用户端和教师端公有的功能模块，后两个功能是普通用户独有的功能。

1. 账号管理

本模块为用户提供登录、注册、修改个人信息（包括修改密码）的功能，用户必须先登录系统，才能进行相关操作。在登录时，用户不仅需要输入正确的账号密码，还必须选择正确的账号类型。登录后，可以修改自己的个人信息，包括姓名、学号、年级、专业、所属学院以及账号密码，修改密码后需要重新登录系统。

1. 论文检测

本模块是整个系统的核心功能，主要包括查看论文检测模版、论文上传、论文检测、查看检测结果、下载检测报告和查看检测历史记录的功能。该模块提供论文模版详细信息的展示功能，方便用户选择合适的论文模版。同时，需要对上传论文进行初步的检查，保证上传的论文格式正确，即上传的文件是DOC或DOCX文件。在网页端上，系统要引导用户输入论文中英文标题，选择合适的论文检测模版后再进行检测。检测完成后，给出用户的检测报告，如果用户检测通过，则提供论文的PDF格式下载。

关于论文检测功能的具体使用流程设计如下：

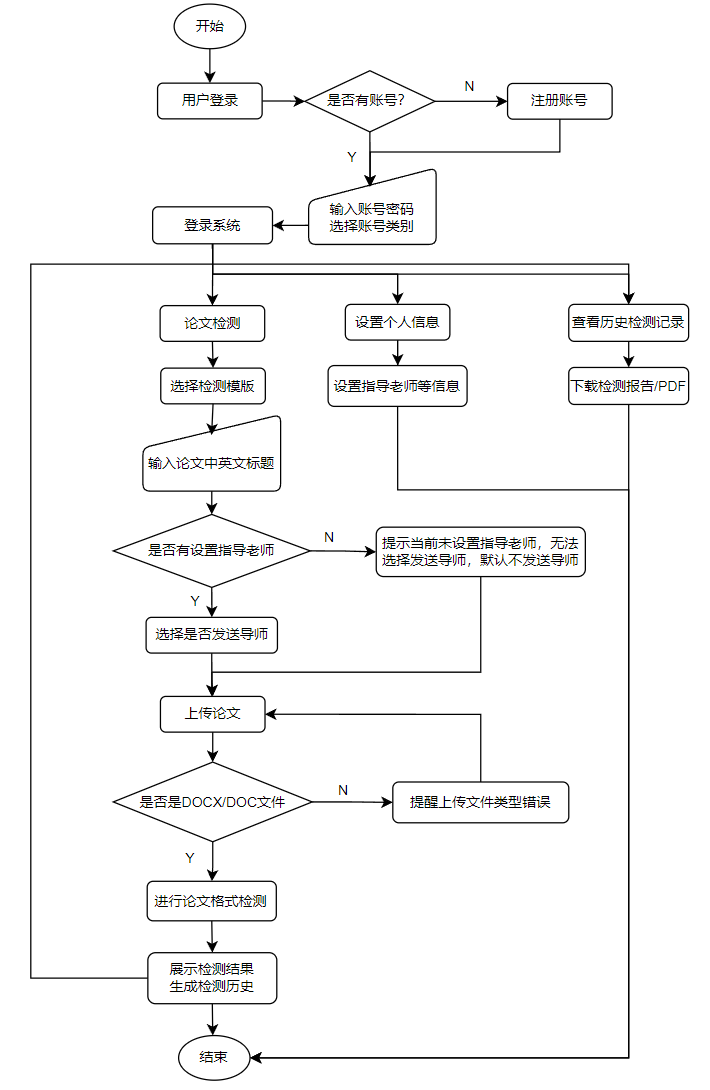


图 5-3 论文检测功能使用流程设计

1. 绑定指导老师

本模块要为用户提供绑定指导老师的功能，首先要为普通用户端加载教师账号信息，为用户提供选择列表。同时，为了方便用户绑定指导老师，还需要提供搜索功能，让用户能通过指导老师的名字快速搜索到目标老师。

1. 发送论文到指导老师

当用户绑定了指导老师后，在检测论文时可以选择将检测结果发送给导师，若本次格式检测通过则会将本次检测记录发送给指导老师，如果本次检测没通过则不会发送。这样能确保发送给导师的论文都是通过了格式检测的论文，让导师可以不必花费时间在检测学生论文格式上，从而有更多的时间查看学生论文在内容上的问题。

### 教师端

教师端主要用于管理论文检测模版，同时能够查看学生提交的论文，下载学生论文和检测报告。与普通用户端相同，教师端同样可以使用论文检测功能。

1. 论文检测模版管理

对于检测论文模版，每个老师可以登录教师端创建、修改、删除自己的论文检测模版。创建模版时会自动为模版生成唯一的模版id，教师用户需要输入论文模版的模版名称和模版的描述信息，并且逐一设置论文各个部分的格式要求，系统要提供默认的格式要求。修改和删除模版时只能修改和删除自己创建的模版，对其他教师创建模版，只能查看具体的模版信息，不能进行修改和删除。每个论文检测模版有两个状态，已上线和未上线，只有上线的论文模版才可以用于格式检测，未上线的模版普通用户无法查看。

1. 查看学生提交论文和下载学生论文及其检测报告

教师可以查看学生提交的论文。可以按名字、日期、学号等信息筛选学生提交的论文。可以下载查看学生的检测报告（可能会有一些系统建议修改的内容），也可以下载学生论文的原件，还可以下载PDF文件。

1. 账号管理

与普通用户端相同，教师端也提供账号管理功能，包括登录、注册、修改个人信息等功能。教师登录后，可以查看和管理自己的个人信息，包括姓名、教师编号、所属学院、描述信息等。

1. 论文检测

与普通用户端相同，教师端也可以进行论文检测，仅仅只是在进行论文检测的时候教师用户不用选择是否发送导师，其余功能包括查看历史记录、下载检测报告等都与普通用户相同。

### 管理员端

管理员端主要用于管理系统账号、批量导入账号、批量导入学生与老师指导关系。管理员可以查看系统内的所有账号，进行创建、修改、删除等操作。系统提供读取excel文件的方式批量导入账号，在excel文件中填入学生姓名、学号、指导老师等信息就可以自动生成对应的学生、老师账号和学生和老师的指导关系。

## 论文检测程序设计

基于第三章对论文检测的需求分析和第四章对DOCX文件底层组织方式的分析，系统设计的论文检测程序的运行流程设计如下。

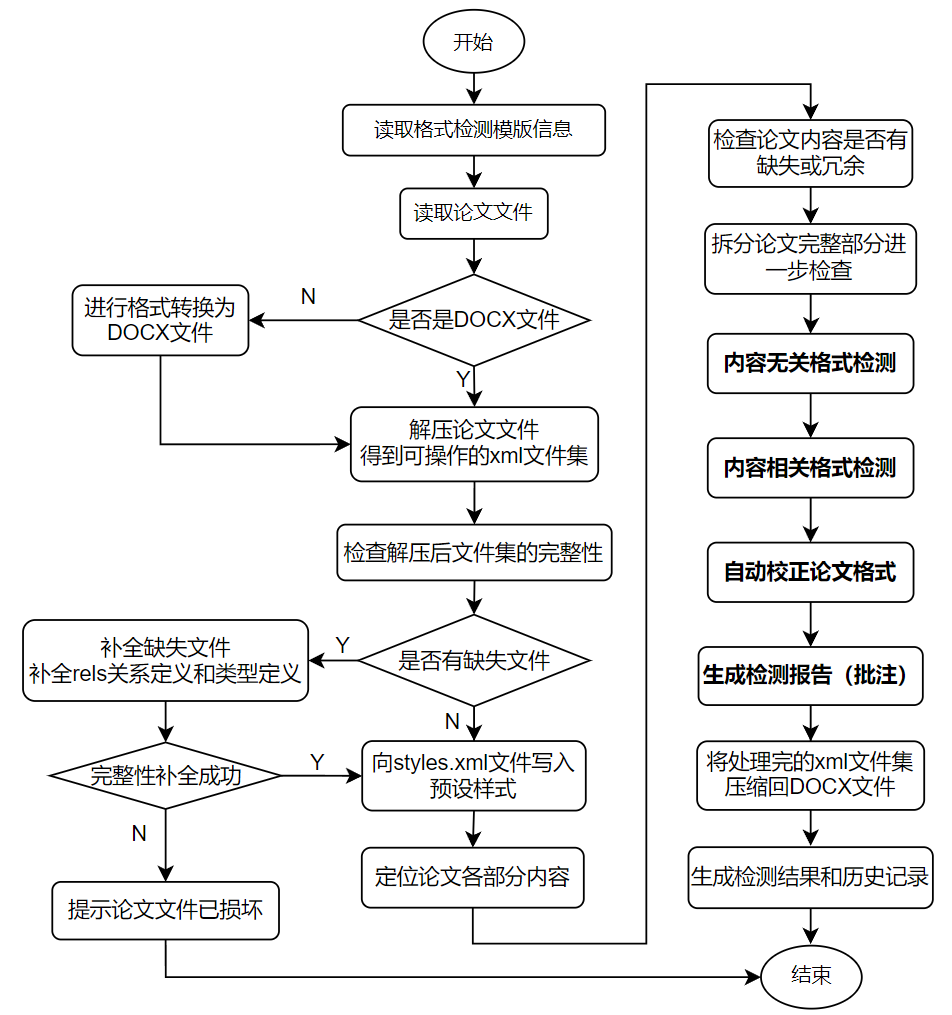


图 5-4 论文检测程序的运行流程设计

如上图所示，系统首先读取格式检测依据的模版文件和要检测的论文文件，然后进行一些预处理，包括格式转换、解压、检测xml文件集是否完整和写入预设样式，随后程序定位论文的各部分内容，即定位论文摘要、目录、正文、致谢等内容，初步检查论文的内容是否有缺失或冗余，对缺失的部分给出提示，对完整的论文部分依次进行内容无关格式检测、内容相关格式检测、自动格式校正和生成检测报告（批注），最后将处理完的xml文件集压缩回DOCX文件，生成本次检测结果和历史记录。

### 内容无关格式检测功能设计

内容无关格式检测不需要系统结合论文内容进行检测，只需要定位到该部分格式控制标签的位置，即可进行格式检测。如图（TODO）所示，本系统对内容无关格式检测主要设计两种方法。

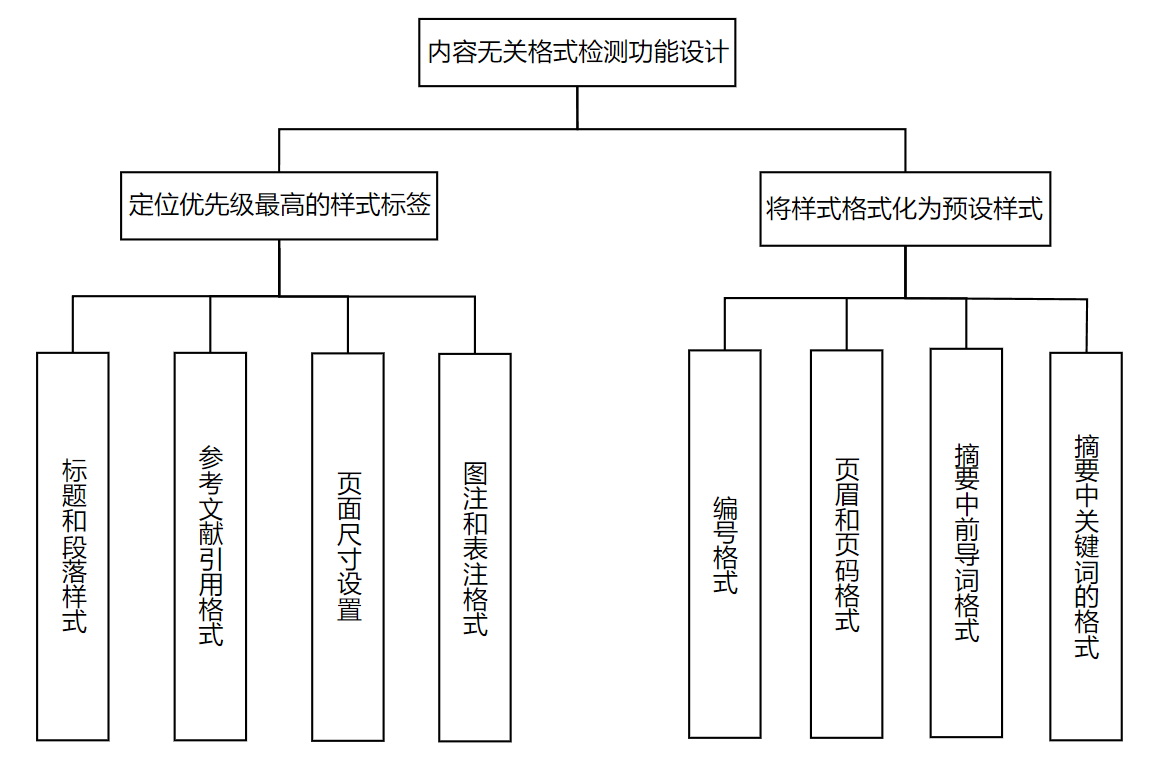


图 5-5 内容无关格式检测功能设计

1. 定位和修改样式标签

根据DOCX文件底层对样式的多层级样式控制，系统定位各层样式控制标签的位置，找出优先级最高的样式控制标签，检测该标签的内容是否符合模版规范。这种方法主要用于标题段落的样式检测、参考文献的引用格式检测、页面尺寸设置检测和图注表注格式检测。

1. 将样式格式为预设样式

对于一些格式比较固定，但底层样式控制标签比较复杂的样式，系统会在定位到样式控制位置后，直接将原有的样式控制标签删除，替换为提前预设好的样式。这样的好处是可以不必分层级找到最高优先级的样式控制标签，也不必深入检测各个样式是否符合规范，可以直接把样式“格式化”为正确的样式，提高了检测的效率。这种方法主要用于编号格式、页眉和页码格式、摘要中前导词格式、关键词的格式。

### 内容相关格式检测功能设计

内容相关格式检测需要系统结合论文的实际内容进行检测，在论文中定位到位置后，读取该部分的内容，根据对内容分析的结果来进行格式检测。如图所示TODO，本系统对内容相关格式检测功能设计有七个部分，主要分为两类，一类只与某部分内容有关，一类与全文内容相关。

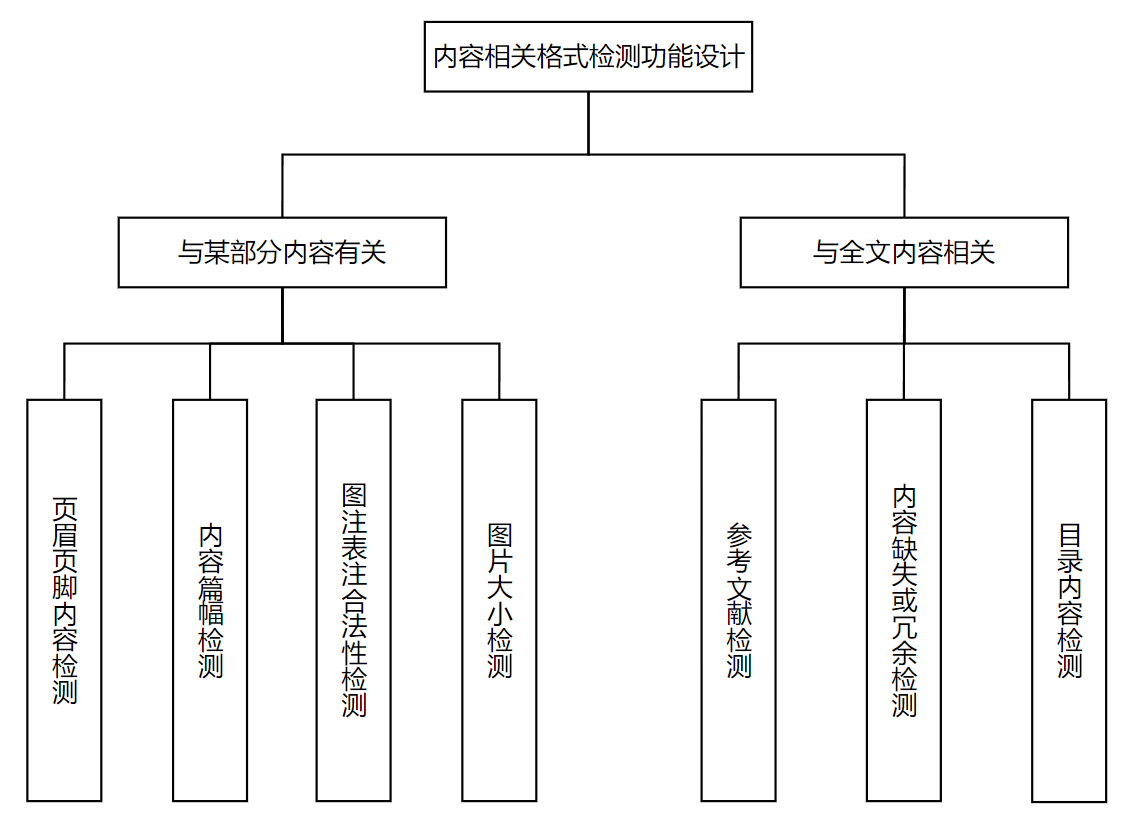


图 5-6 内容相关格式检测功能设计

1. 与某部分内容相关

该部分的检测需要先定位到具体内容的位置，读取内容后进行检测。

①页眉页脚内容系统会读取每一节的页眉页脚中的内容，根据内容进行检测。

②内容篇幅检测会读取论文各部分的内容，统计字数（或单词数）判断篇幅是否符合要求。

③图注表注合法性检测会读取每个图片和表格向上下文段的内容，保证每个图片都有图注，每个表格都有表注，同时读取每个图注表注上下文段的内容，保证每个图注表注都是合法的图注表注，即每个图注表注都有对应的图或表；

④图片大小检测会读取控制图片的标签，根据标签内控制图片尺寸的内容检测图片尺寸是否符合要求。

1. 与全文内容相关

该部分的检测需要对全论文的内容进行遍历。

①参考文献的检测有两方面，一方面是需要用正则表达式保证每条参考文献的内容按照标准规范组织，同时统计参考文献的数量，保证数量符合要求；另一方面是需要检测论文全文，保证所有参考文献都是有效的参考文献，没有未被引用的参考文献，也没有引用了不存在的参考文献。

②内容缺失和冗余检测是在系统对论文各部分进行初步定位后，判断是否有缺少和冗余的部分。

③目录内容检测需要系统遍历全文内容，读取论文正文内容中的一级、二级和三级标题，确保这些标题内容都在论文的目录中。

### 报告生成与自动校正功能设计

系统报告是以批注形式给出，在检测过程中会存储每个需要插入批注的位置和批注的内容，在检测完成后，根据这些位置信息和批注内容，逐一生成和插入到论文中去。结合DOCX批注的底层组织方式，系统提供七种批注插入方式，包括文段首部插入、文段末尾插入、整段选中插入、文块首部插入、文块末尾插入、全文末尾插入、范围选中插入。

对于自动校正功能，系统设计提供对所有内容无关格式的自动校正和内容相关格式中页眉页脚部分的自动校正，设计以下面两种方式实现。

1. 修改标签内容实现

与内容无关格式检测中“定位和修改样式标签”方法一样，系统找到最高优先级的样式控制标签，如果检测到格式错误，则直接修改标签的内容实现对格式的自动校正。

1. 格式化标签内容实现

与内容无关格式检测中“将样式格式为预设样式”方法一样，在需要矫正的地方直接用预设的内容替换掉原有的内容，实现对特定内容的自动校正。

# 系统实现

交互需求：具体有哪些页面。

使用流程图。

具体界面展示

# 系统测试-----

# 结论

# 致谢

感谢党和国家。

感谢指导老师，和培育过我的老师。

感谢父母

感谢我自己

感谢我的导师XXX老师，谢谢她对我的悉心指导。她无私的关爱和严谨的治学态度，将激励我不断的进取，走好以后的道路。其次，还要感谢在这四年的学习中教过我的所有老师们，谢谢他们传授给了我知识。我的同学XXX，在写作的过程中给我提供了一些宝贵的资料和建议，在此一并感谢！



# 参考文献

1. 吴海波,陈拾菊.本科毕业论文质量评价标准研究[J].绍兴文理学院学报,2023,43(01):105-112+121.DOI:10.16169/j.issn.1008-293x.j.2023.04.017.
2. 祝艳.试析学位论文质量的内涵与标准[J].文教资料,2018,(07):138-139.
3. 刘树林.高校本科毕业生学位论文格式存在的问题与解决路径[J].中国教育技术装备,2023,(12):121-124.
4. 常兴治,虞菊花,阚琦.高职毕业论文文档格式规范化的探究与实践[J].常州信息职业技术学院学报,2020,19(05):51-53.
5. European Computer Manufactures Association. ECMA-376: Office Open XML File Format-Fundamentals And Markup Language Reference[EB/OL]. Brussels: ECMA, 2021 [2024-04-02]. URL: https://ecma-international.org/ publications-and-standards/standards/ecma-376/
6. Fu Z, Sun X, Liu Y, et al. Text split‐based steganography in OOXML format documents for covert communication[J]. Security and Communication Networks, 2012, 5(9): 957-968.
7. 张印.基于VSTO的高校毕业论文格式化处理软件的设计与实现[D].北京邮电大学,2012.
8. 陈渊博.学位论文格式检测系统设计与实现[D].大连理工大学,2017.
9. 徐俊.毕业论文格式检测与校正系统研究与实现[D].重庆邮电大学,2022.DOI:10.27675/d.cnki.gcydx.2022.001234.
10. 张巍巍.学位论文格式规范性自动检测系统设计与实现[D].大连理工大学,2019.DOI:10.26991/d.cnki.gdllu.2019.002102.
11. 袁敏.学术论文格式检查和内容校对的研究[D].北京交通大学,2019.
12. 曹敏.GB/T 7714-2015《信息与文献参考文献著录规则》标准解析[J].科技与出版,2015,(09):41-44.DOI:10.16510/j.cnki.kjycb.2015.09.013.