附件2：



**内蒙古师范大学计算机科学技术学院**

**毕业设计（论文）开题报告**

**题 目： 基于OpenXML的文档格式解析程序**

**专 业 计算机科学与技术**

**姓 名 齐家宝**

**学 号 20201103035**

**指导教师 翟晔**

**日 期**

**计算机科学技术学院制**

**说 明**

**一、开题报告主要内容**

1．课题来源及研究的目的和意义

课题来源：随着计算机领域的飞速发展以及信息技术的普及。各项工作都离不开计算机的参与，办公领域office的使用更是随处可见。对于各大高校来说本科生和研究生的论文撰写都是在电子文档上完成的。学位论文的撰写是特别重要的教学计划环节之一，也是教学计划组成的重要部分，是检验本科生和研究生的学习成果和学习质量的核心部分。[1]每年的本科生和研究生的毕业生数量繁多，因而相应的学位论文数量庞大，随之带来的问题显而易见，论文格式的检查工作是巨量的，繁琐的，也是枯燥的。[2]学生需要花费大量的时间精力去完成论文格式的检查和校验。[3]对于论文撰写的规则与模板格式必须要符合学校规定的要求，这导致大部分学生在论文格式的检查和修改上投入了大量的时间精力，如果把这部分时间精力花费在论文的内容撰写和学术研究上显然更合适。[4][5]论文格式也会在一定的程度上影响学生的学位论文通过率，所以指导老师也需要大量的时间和精力去检查学生论文格式的合规性。因为每个人的时间精力是有限的，再加上检查论文格式是否符合标准浪费了大量时间精力，这将导致指导老师无法更多的去关注论文内容，无法更多的去帮助学生检验论文内容的科学性，独特性，有效性等。[1]长期以来会出现重视论文格式，不重视论文内容，这种不清楚孰轻孰重颠倒轻重的情况。[1][4]论文格式的解析工作重复性巨大，且人工效率极低，反而计算机是对此类工作是很擅长的也是很适合的。在过去（Microsoft Office 2000系列以前）office软件文件还是OLE格式（Unicode 编码）。从Microsoft Office 2000 系列开始微软公司的office开始支持XML.。到Microsoft Office 2007时微软公司已经彻底用XML替代了原先office的二进制文件存储方式。[6]Microsoft Office 2007以后（包括Microsoft Office 2007）所使用的XML格式也被命名为OpenXML。如今是国际文档格式标准了。[7]XML格式的文档文件特别适合计算机解析。

研究的目的和意义：如果能让计算机完成文档格式的解析工作，这将能大大提高论文文档的格式校验工作的效率与准确性，减少重复性工作，提升整体效率。能有效减轻学生和指导老师在论文格式检查和修改上的时间精力投入，让学生和指导老师更多的关注论文内容，从而提升论文内容的质量。这也将使得论文撰写更加规范，符合学校规定的要求，从而提高学位论文的通过率。在论文格式解析工作中计算机的运用也能促进计算机技术在学位论文撰写过程中的应用，提高计算机技术的使用效率，推进信息技术在教育领域的普及和应用。同时，也可以为其他领域的文档格式解析提供参考和借鉴，具有较广泛的应用前景和较高的研究价值。

2．国内外在该方向的研究现状及分析

目前主流的办公软件是微软公司的Microsoft 365 （Office），其中文字处理软件是Word，可以说微软的office主导着办公软件的领域。微软公司开发的Word软件也是占据着文字处理软件中的核心位置。但是微软公司对于Word软件的开发文档或者API的共开较少。利用程序代码操作Word文档的技术公布也相对较少。微软公开了Office Open XML标准，又给提供了Open XML SDK。但是此SDK提供微软自家语言C#的方法和接口。C/C++.Java等语言调用此SDK并不容易，也SDK本身没有提供相关方法和接口。关于C++操作Word可值得一提的是微软提供了COM API可以借助COM API用C++编程实现Word的自动化。但是COM API需要依赖office组件，在一台没有office的计算机上借助COM API开发的程序是无法正常运行的，这也是致命的问题。

目前，国内外许多研究和开发机构都在致力开发文档格式解析程序的研究和开发。例如，国内的一些高校和科研机构正在进行相关项目的研发，如北京理工大学的“毕业设计论文格式检查系统”等，但是此类相似研究目前还是较为匮乏。

Word文档格式检查校验需求市场份额并不是很乐观。再加上企业和各个科技公司对于Excel（电子表格）的各类操作需求要远大于Word文档格式校验的需求。所以在Excel表格的处理技术上更用心导致Excel表格的操作处理手段和方法技术要比Word文档丰富的多，Word在这一方面属于在匮乏的状态，再一方面，开源社区对于Word格式处理的重视度并不高。再加上Word格式处理的开发文档与技术文档和各类研究文献匮乏导致开发关于Word的格式解析和处理程序较为困难所以Word格式处理问题上研究并现状不容乐观。[4][8]

现有的一些Word格式解析检查程序的程序大部分基于微软的.NET平台的VBA技术或者VSTO（全称Visual Studio Toolfor Office,是.net平台下的Office开发技术）技术。VBA技术缺陷也是很明显，无法达到全自动化，程序操作期间还需要大量人工辅助完成工作。利用VBA技术完成的Word格式检查程序要依靠大量的人工操作，这显然不是我们想要的结果。[8]替代VBA技术的方法之一就是用微软新提供的office开发技术VSTO进行office处理的程序开发。但是VSTO技术开发对于平台的依赖性特别大，还需要特定的开发工具，VSTO技术需要依靠微软提供的IDE（集成开发环境）：Visual Studio。Visual Studio是windows平台的开发工具，在其他平台也是不方便使用。[9]

这一现状维持到Microsoft Office 2007问世以后，微软提出了Open XML格式随后ECMA-376 标准问世。[7]并且微软公开了Open XML标准相关的文档和说明。其中包含针对Word文档处理的相关参考手册。现如今对于word的操作我们可以依靠xml处理技术完成相关操作。[10][11]

3．主要研究内容

OpenXML文档结构：了解OpenXML文档的结构和组成，包括文档类型、文档元素、命名空间、属性等。

XML解析：使用XML解析器对OpenXML文档进行解析，将其转换为内存中的树形结构，便于后续处理。

文档组件的提取：根据OpenXML规范，将文档分解为各个组件，如标题、段落、表格、图片等，并提取其属性和内容以及研究文档格式的特点、规范和标准，和相关的解析工具和库的使用。

文本提取和识别：对于文本内容，需要进行文本提取和识别，以便于后续的文本处理和分析。

数据操作和处理：对提取的文档组件和文本内容进行各种操作和处理，如格式化、排序、分类、搜索等。

文档生成：将处理后的数据重新生成OpenXML文档，保证格式的正确性和美观性。

Xerces-C++的使用和研究：Xerces-C++是一个用于处理XML和XML-based语言的开源C++库。它提供了一组API和工具，用于解析、验证、操作和生成XML文档。研究Xerces-C++库的原理、机制和使用方法，是实现文档格式解析程序的关键。

数据结构和算法的应用：文档格式解析涉及到大量的数据结构和算法的应用，如树形结构、递归、解析树等。研究这些数据结构和算法，并将其应用到文档格式解析中，可以提高程序的性能和效率。

4. 研究方案

研究OpenXML和Xerces-C++库：深入了解OpenXML和Xerces-C++库的API和使用方法，掌握如何使用这些库来读取、解析和操作文档格式。

解析归纳整理OpenXML标签：OpenXML标准内容庞大，标签种类繁多，需要进行归纳总结，了解每个标签的含义。

设计解析程序架构：根据需求和分析文档格式，设计解析程序的架构，包括输入、处理和输出等模块，并确定使用哪些库和工具来协助实现解析程序。

实现解析程序的核心功能：实现解析程序的核心功能，包括读取文档、解析文档、访问文档元素和属性等。具体来说，可以使用Xerces-C++库提供的API，读取和解析XML文档，使用OpenXML库提供的API，访问和操作文档中的元素和属性。

实现解析程序扩展功能：根据需求和分析文档格式，实现一些扩展功能，如搜索、过滤、修改等。可以使用OpenXML和Xerces-C++库提供的API，实现这些功能。

测试和调试：对解析程序进行测试和调试，以确保程序能够正确地读取和解析各种文档格式。可以使用各种测试用例和工具来验证解析程序的正确性和性能。

优化和改进：根据测试和调试结果，对程序进行优化和改进，以提高程序的性能和可靠性。例如，使用缓存技术来提高程序的速度，使用多线程技术来提高程序并发处理能力等。

文档化：对解析程序进行文档化，以方便后续维护和扩展。文档应该包括程序的设计思路、API使用方法、测试结果等。

5．进度安排，预期达到的目标

第一阶段：

研究OpenXML和Xerces-C++库的API和使用方法。

了解OpenXML标准的基本概念和XML文档结构。

使用Xerces-C++库读取和解析简单的XML文档。

使用Xerces-C++库读取和操作Word等格式的文档。

编写简单的测试程序来验证库的使用方法和功能。

第二阶段：

设计解析程序的的结构和模块划分。

实现解析程序的核心功能，包括读取文档、解析文档、访问文档元素和属性等。

能够解析复杂的文档格式，包括模板Word文件，目标Word文件等。

编写详细的测试用例和测试程序，对解析程序进行全面的测试和调试。

第三阶段：

实现解析程序的扩展功能，如搜索、过滤、修改等。

能够根据需求进行定制和扩展，增加程序的功能和灵活性。

对解析程序进行优化，提高程序的性能和响应速度。

第四阶段：

对解析程序进行全面的测试和调试，确保程序的正确性和稳定性。

优化程序性能，提高程序的效率和响应速度。

解决程序中出现的问题和bug。

第五阶段：

对解析程序进行文档化，编写清晰、易懂的文档。

整理程序代码和文档，以便后续维护和扩展。

提供用户使用文档和示例程序，方便用户使用和理解程序。

6．课题已具备和所需的条件、经费

一台或更多的计算机，能够运行Windows操作系统和相关的开发工具。

安装有Visual Studio等集成开发环境（IDE）。

安装有OpenXML和Xerces-C++库的相关开发工具和头文件。

7．研究过程中可能遇到的困难和问题，解决的措施

技术难度高：文档格式解析涉及到的技术和知识领域比较广泛，包括XML、DOM、DTD、XSD、OpenXML、Xerces-C++等，可能会遇到技术难度较高的问题。解决措施可以是加强相关技术的学习和研究，查找相关的技术文档和资料，以及向相关技术人员寻求帮助。

复杂性大：文档格式解析程序需要处理复杂的的数据结构和算法，可能涉及到大量的计算和存储资源，程序复杂度也较高。解决措施可以是采用一些优秀的设计模式和算法，对程序进行模块化和分解，避免过于复杂的实现。

对于OpenXML格式的文档，可能需要处理复杂的的数据结构和数据类型，例如图表、图片、公式等。解决措施可以是深入研究OpenXML规范，了解不同类型的数据如何表示和解析。

对于Xerces-C++解析库，可能需要调整解析参数以适应不同的文档格式。解决措施可以是深入研究Xerces-C++解析库的文档格式，了解如何调整解析参数以适应不同的文档格式。

8．主要参考文献

**二、开题报告要求**

1．开题报告的字数应在2000字以上。

2．参考文献的要求：

（1）理工类论文的参考文献一般为10-15篇，其中学术期刊类文献不少于7篇，外文文献不少于3篇（特殊专业可酌情确定明确要求，并报教务处备案）；文科、管理类论文，参考文献一般为15-20篇，其中学术期刊类文献不少于12篇，外文文献不少于3篇。近五年的文献数不应少于总数的1/3，应有近两年的参考文献。教材、产品说明书、国家标准、未公开发表的研究报告不宜作为参考资料。

（2）参考文献按在开题报告中出现的次序列出。

（3）参考文献书写顺序：序号 作者.文章名.学术刊物名.年，卷（期）：引用起止页。

（正文 宋体小4号字，行距1.25倍，段前0行，段后0行）

1. 王杏龙 & 毛永江.(2008).本科毕业论文质量下降的原因及对策. 高等农业教育(09),69-71.
2. 喻健 & 朱三元.(2016).基于Entity Framework和easyUI的毕业论文管理系统. 软件工程(12),40-42.
3. 阚运奇.(2012).毕业论文格式检测系统的设计与研究. 无线互联科技(11),17.
4. 徐东风,彭红星 & 廖俊杰.(2010).基于Java的文档格式检查技术的研究及其应用. *计算机工程与设计*(19),4309-4311+4315. doi:10.16208/j.issn1000-7024.2010.19.018.
5. 褚佳强.(2009).计算机辅助教学管理——本科毕业论文规范化排版. 科技资讯(28),232. doi:10.16661/j.cnki.1672-3791.2009.28.128.
6. 赵振洲 & 乔明秋.(2013).基于OXML的Office文档结构分析与取证调查. 计算机安全(08),44-46.
7. (2008).微软文档格式获批为ISO/IEC标准. 每周电脑报(11),41.
8. 孙晓伟,毕晓清,吴洪越 & 赵卫东.(2013).基于DOM和VBA的Word文档数据交互式抽取. 计算机应用与软件(02),113-115.
9. 刘艳平.(2016).基于VSTO的word文档格式检查插件的开发与研究. 电脑知识与技术(26),106-108. doi:10.14004/j.cnki.ckt.2016.3435.
10. 李琳.(2003).浅谈XML及其在Microsoft Office中的应用. 衡阳师范学院学报(自然科学)(03),69-72. doi:10.13914/j.cnki.cn43-1453/z.2003.03.020.
11. 文档格式标准攻防战尘埃落定 微软过关各方积极应对[J]. 中国计算机报,2008, (12).