基于MathJax的论文解析和Web展示

摘 要

针对目前基于LaTex排版的论文在网页上显示多采用文章内容扫描的方式所导致的文章内容模糊且难以阅读的问题，开发了一种面向互联网用户的论文解析与展示的服务。在充分考虑了LaTex的排版方式和数学公式特点的基础上，通过对LaTex排版方式的分析和总结，并利用MathJax网页渲染引擎的优秀实现效果，为服务进一步的实现打下基础。通过MVC模型（javabean+servlet+jsp）建立了该web服务的基本框架。通过各子功能的测试表明，该服务完美实现了LaTex论文解析和web展示的预期，有利于解决目前论文网页展示的弊端。

首先本文介绍了LaTex的基本信息，对目前主流网页处理LaTex论文的方式进行分析，指出其存在的优势与不足，并在此背景下提出一种解决方案；随后介绍服务的实现过程与技术难点，最后通过前后处理的结果对比总结该服务的必要性和先进性。

关键词：LaTex；web展示；公式渲染；

# 概述

## 研究背景与意义

信息时代的到来，为人们信息的交流提供了便利，虽然互联网上的信息量正以几何级数的方式在增长，但是由于承载这些信息的数据在生产过程中的组织和存储方式不一样，导致这些信息很难被不同的平台共享。对于学者来说最直观的莫过于论文的阅读。现在流行的LaTex作为一个强大的论文编写工具，近年来其优势明显、上手容易的特点得到了论文作者的极大关注，目前用户大多通过互联网查阅论文，目前主流网站上搜索得到多为作者上传的PDF格式的论文，用户通过浏览器下载文件或者在线阅读通过扫描得到的对应文本，但在线阅读的论文因为其模糊不易阅读的特点使读者烦恼不堪。对用户和作者造成了很大的不便因素，与此同时因为html格式的论文的便利性与显示清晰的特点受到许多学者的追捧。为了保证论文能更方便的交流和发布，必须进一步完善解析LaTex的源文件的功能，以期达到最好的效果。同时PDF作为论文的标准格式其论文内容无法选中、标注，且PDF格式的论文中数学公式一般通过插入图片的方式来显示，易导致论文的排版混乱的问题。为了解决这种种不便，开发一款LaTex的web显示服务是所有读者的迫切需要。目前在LaTex的web显示中数学公式一般是通过由TEX向MathML转换后得到，虽然一些主流浏览器并不支持MathML标准如IE需要安装MathPlayer插件才得以使用，这增加了用户的不便，但是其优势仍然不容忽视。

## 国内外研究现状

在LaTex的解析和Web显示技术的发展过程中，针对如何高效、准确的辨别论文内容的问题上，国内外相关领域的研究者提出了一些解析算法，很多高校以及研究机构也在此基础上实现了很多优秀的Web显示系统。参见编译原理[1]过程可分为三步，第一步：词法分析，第二步：语法分析，第三步：语义分析。三个步骤缺一不可，但是由于LaTex语法的复杂性解析过程十分繁琐，实际设计过程中不得不牺牲一定的效率以保证其准确的结果，而由于语法中标签的复杂嵌套关系，单纯的分析标签代表的意义显的毫无意义，要通过大量的论文源文件的测试，丰富解析库来达到准确的目的。

国内学者对LaTex的使用相对起步较晚，但是后来者居上，如今在国内的网站上可以见到大量LaTex格式的论文，同时在贴吧内也能见到许多人在交流LaTex使用经验，对于新手来说LaTex原先由于缺乏教程入门极难编写出来的论文格式混乱，一些高校研究机构编写了一些LaTex的模版，使用者只需要在对应的位置填入文章的对应内容就可以完成一篇论文，对于论文的初写者节约了大量的排版时间，同时论文格式的统一与建立，极大地方便了信息的交流。

LaTex技术发展至今，虽然没有经历太长的历史，但是在国内外研究人员的不断探索下，这项技术正在不断向前发展并一步步趋向成熟，与此同时也面临着一些新的挑战，很多问题急需进一步的研究。最主要的几个问题是：首先是LaTex的标准统一，纵观信息技术发展的历史长河，新的统一的标准需要在一个特定的时机才能由相关的国际组织提出来，新标准的普及也需要很长的一段时间。其次是很多用户因word与LaTex的便利性而争论，LaTex需要通过各种标签来规定内容的排版，而word需要手动调节字体的大小等因素，两者互有利弊但是便利性的评判具有较高的主观性。

## 主要研究内容

本项目以LaTex的标签解析对象，分析总结现有解决方法的优势和缺陷，在继承其优点的同时对准确性、高效性等方面作进一步的改进，基于MVC模型，优化数据库表项及后台数据处理逻辑来提高处理的效率；优化界面美工，使用AJAX简化用户的操作，增强网页与服务器之间的交互以提升用户体验。本项目具体实现的功能如下：一是Latex源文件的检测上传及保存至服务器；二是LaTex源文件通过词法分析，语法分析，语义分析三个步骤来解析（包括题目、摘要、关键词、内容、数学公式、参考文献等），并将解析之后得到的数据保存成jason数组返回给网页；三是解析后内容的按原论文的排版显示，尤其是数学公式需要通过Mathjax引擎渲染；四是后台管理界面的设计与实现，用户交互界面的设计与实现。

研究方法：该设计通过面向对象的开发方法，利用面向对象的概念和方法为软件需求建立模型，进行项目设计。

技术路线：

1）按照系统可行性进行模块的划分，完成整体实现思路。  
 2）完善数据库设计（采用SQL数据库进行开发）。  
 3）主要采用MVC（JSP+Servlet+Javabean）设计模式来设计。  
 4）JSP及CSS实现提升用户体验。  
 5）JQuery框架实现用户体验的进一步提升。  
 6）前后台各层间的整合与协调。

1. 相关理论

# LaTex

## LaTex定义

LaTeX是当今世界上最流行和使用最为广泛的TeX宏集，由美国计算机学家莱斯利·兰伯特开发，它避免了没有排版和程序设计知识的使用者花费大量的时间去学习排版，适用于各种高印刷质量和复杂的的文档，由最原始的TEX格式不断发展丰富，随后产生了Plain TeX与低层次的TEX相呼应，而如今的LaTex是以Plain Tex为基础，加入了很多的新功能极大简化了使用者的操作。

|  |
| --- |
| [1]王福. 基于Web的编译原理学习支撑系统的设计与实现[D].中南大学,2014. |