**高校论文查重系统**

**可行性分析**

王显龙 15301051

林泽鑫 15301125

刘华秋 15301126

骆人杰 15301130

王瑞华 15301136

王婉铃 15301158

## 项目背景及简介

## 1.项目背景

近年来各大高等院校愈加重视学术打假。每年有大量的本科应届生、研究生提交自己的学术论文，其作为学术造假的一个主要源头应被严格把关。但现有的公共论文查重平台，如CNKI、万方等，虽然收录了大量期刊、会议论文，但对高校本科生、研究生所发表的论文收录并不全面，故出现了对高校内本科生、研究生论文搭建论文库进行查重的需求。高校论文查重系统提供了这样的平台，既为各高校提供定制化论文库的论文查重服务，更为各高校提供了共享私有论文数据库，形成高校论文联盟、提高查重效果的平台。

## 2.项目功能简介

高校论文查重系统主要分为专为高校内网使用的论文查重系统和架设在公网的论文查重系统两种形式。本论文查重系统主要功能如下：

1. 论文库论文的更新、删除、列表查看

即每个用户可以维护自己的论文查重库，对库中的论文进行增删查操作

1. 论文批量查重

即用户可以批量上传待查重论文，将上传的论文与论文库中的论文进行查重

1. 多篇论文互相抄袭检查

即用户可以批量上传论文，并检测该批论文之间有没有互相抄袭的情况

1. 论文查重报告批量下载

即用户进行完论文查重操作以后，系统会为每篇论文生成的查重报告，用户可以进行批量下载

1. 发送论文库互相共享邀请/接受论文库共享邀请

即各高校之间可以发送 私有论文库互相共享的请求，对方同意后，双方之后进行的查重将会与双方的私有库同时进行比对

## 项目技术可行性

## 前端设计技术可行性

前端主要使用BootStrap框架进行页面设计，主要技术包括HTML、CSS、JavaScript。前端设计主要包括论文库增删查页面、论文查重功能页面、查重结果展示页面。主要难点在于查重结果展示形式的设计，可以通过表格、图片等形式展示结果。在公网环境中的论文查重系统，还需要设计登录注册、高校审核和认证高校搜索页面，技术上均可通过BootStrap框架得以实现。

## 2.服务器端技术可行性

服务器端主要使用Spring + SpringMVC + MyBatis框架进行逻辑处理，包括登录注册、认证高校搜索、添加论文到论文库、删除论文、论文格式转换和论文查重功能。其中论文格式转换功能涉及到pdf、word、txt等格式，需要对这些相关格式的转换工具进行研究，设计转换逻辑，并对论文中不需要比对数据（如承诺书、目录等内容）的删除逻辑进行设计。已知可使用的格式转换工具有Acrobat接口、Office接口、PdfBox、pypdf2等。论文查重算法上，需要考虑查重算法的时间复杂度，同时需要考虑语句重复判定的逻辑，需要对现有查重逻辑、自然语言处理等相关内容进行研究。在自然语言处理上，有word2vec模型，可以将文字转换为向量，进行相关计算。

服务器端功能开发有一定难度，但在技术方面都有相应的解决方案，需要花费一定时间进行学习和设计。

## 项目数据源取得可行性

1、项目前期可通过校园图书馆从学校购买的数据库中批量下载部分论文用于数据分析。

2、小组成员有“爱学术”网站会员，可通过“爱学术”网站进行批量论文下载。

3、项目开展以后会与软件学院研究生院相关负责老师洽谈，争取相关数据。

## 项目市场可行性

项目开题前进行过少量的市场调研，通过对北京交通大学若干学院研究生院的相关负责老师咨询，我们获悉现有传统论文查重系统有以下不足：

1、传统的权威论文查重系统（如万方、CNKI等）只具备在其自有的论文库中查重的能力，不具备同一批次论文间的查重能力，即：如果同一批即将毕业的学生之间论文存在互相抄袭的情况，传统的论文查重系统是无法检测的。

2、传统的论文查重系统的查重论文库中收录的论文，其中大部分是发表在期刊、会议、报纸上的论文。各高校毕业论文和学位论文的收录量远远不够。学校有很多自有的、没有被万方/CNKI等权威机构收录的论文。

3、在有些特定情况下，未毕业学生的论文为防止论文泄露不能上传到互联网上，此时不能使用互联网上的查重系统。如果没有内网查重系统，将无法进行论文查重。

而我们开发的系统区别于传统权威查重机构，面向各个高校，让各高校可以定制自己的查重论文库，且可以进行论文库互相分享。从一下三个方面解决了现有传统系统的痛点：

1. 提供同一批待查论文间的自查重功能。
2. 采用各高校的私有论文库，且各高校之间可以互相共享私有论文库，完美弥补了传统论文查重的此项不足。
3. 依赖于学校的私有论文库，可以部署在学校的内网服务器中，杜绝论文泄露问题。

## 项目时间可行性

项目中主要难点在于论文格式转换、查重算法设计和查重结果展示的设计，这也是项目的核心功能。项目的时间主要花费在前端界面设计、查重算法设计、格式转换工具的学习和使用上，可以通过合理的分工和合理的时间规划，在课时内完成项目的开发。