**高校论文查重系统**

**开题报告**

王显龙 15301051

林泽鑫 15301125

刘华秋 15301126

骆人杰 15301130

王瑞华 15301136

王婉铃 15301158

项目概述：

## 项目来源及背景

学术造假行为已成为当今社会关注的热点问题，尤其是2017年的学术打假热，更是将学术造假推向了风口浪尖。各大高等院校每年有大量的应届生、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更应对这些论文严格把关。高校论文查重系统提供了一个很好的平台，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。

## 项目与实践的关系

积极响应国家《关于加快建立国家科技报告制度的指导意见》。根据意见，对科技报告存在抄袭、数据弄虚作假等学术不端行为的，纳入项目负责人和项目承担单位的科研信用记录并依据相关规定向社会公布。

## 项目开发意义

高校论文查重系统依托现代化的信息技术，建立准确、高效的抄袭判定模型；依托互网技术建立高校私有论文库双向分享平台。其意义主要有以下五个方面：

1、积极响应了国家《关于加快建立国家科技报告制度的指导意见》及其它学术打假政策。

2、有效打击投机取巧的假学术人，净化高校学术空间。

3、保护论文/科研成果原作者的权益。

4、降低抄袭论文的产出，提高高校论文的平均质量。

5、促进各高校间私有论文库的共享，提高全网查重能力。

## 必要性与现实价值

通过与部分学院研究生院院长的沟通，我们了解到了现有传统论文查重系统的一些不足。并将我们要做的高校论文查重系统的必要性与现实价值总结如下：

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。而我们开发的查重系统将会支持此功能。

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。我们开发的查重系统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、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，而我们开发的查重系统，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。

## 项目主要设计内容

高校论文查重系统项目主要分为两个部分：一是服务器端开发，二是web端开发。如果后期开发时间条件允许，还将开发PC客户端。服务器端与web端、服务器端与PC客户端之间通过Json机型消息同步和传输。服务器开发使用java和C++，主要功能是对用户提供登陆/注册的后台功能以及相关数据的存储，对上传的论文进行存储和解析，以及查重算法的实现。web客户端开发使用html/css/js，主要功能是web端的界面、服务器内容展示和交互、论文上传与查重功能的界面交互。PC客户端使用C#，主要功能同web端。

## 项目重难点

本项目的重点在于论文查重系统的后台逻辑设计和查重结果的展示。服务器端需要根据功能模块需求设计合适的数据库结构，选用合理的应用框架，进行抄袭判定模型的建模、查重算法的设计和pdf/word论文解析等。Web端和PC客户端需要完成各种界面的设计及联调跟服务器的功能接口。完成单元测试、集成测试、系统测试。

# 设计方案：

## 基础理论

JAVA

HTML + CSS + JavaScript

C++

C#

MYSQL

JSON

HTTP协议

GIT 版本控制工具

Eclipse的使用

Spring + SpringMVC + MyBatis框架

BootStrap框架

测试方法、理论

UI交互设计、美工图片设计

项目管理基础理论

## 基本功能设计

为便于后续的表达，我们定义以下专用名词。

1、纵向查重：简称查重，将待查论文与论文库中所有的论文进行比对，找出待查论文中所有与论文库中论文相同（或抄袭自论文库中论文）的部分并标记，计算待查论文的重复率。

2、横向查重：检测用户上传的一批待查论文之间是否存在互相抄袭、互相重复。

3、待查论文：将要使用论文查重系统进行查重的论文。

4、论文库：论文库中存放的论文都是已经取得学位的学生的最终版论文。论文查重系统需要一个论文库才能对待查论文进行查重。在查重时，需要将所有的待查论文和论文库中的论文进行比对。

5、私有论文库：各个高校内部自有的、没有完全对公开的论文库。

6、查重报告：对论文进行查重后得到的报告，里面包含了该论文的总体重率、重复来源、被检测出的与论文库中论文重复的句子。

本高校论文查重系统基于B/S架构。前端预期使用Bootstrap框架进行设计，后台逻辑层主要使用JAVA SSM（Spring + SpringMVC + MyBatis）框架进行设计，部分算法使用C++进行设计。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。

2、纵向批量查重（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）。

3、横向查重（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）。

4、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。

5、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

除了上述主要功能以外，为了保证上述功能的正常运转，系统还有登陆、注册、高校身份认证等功能。高校论文查重系统主要功能对应的使用用例如图1所示。

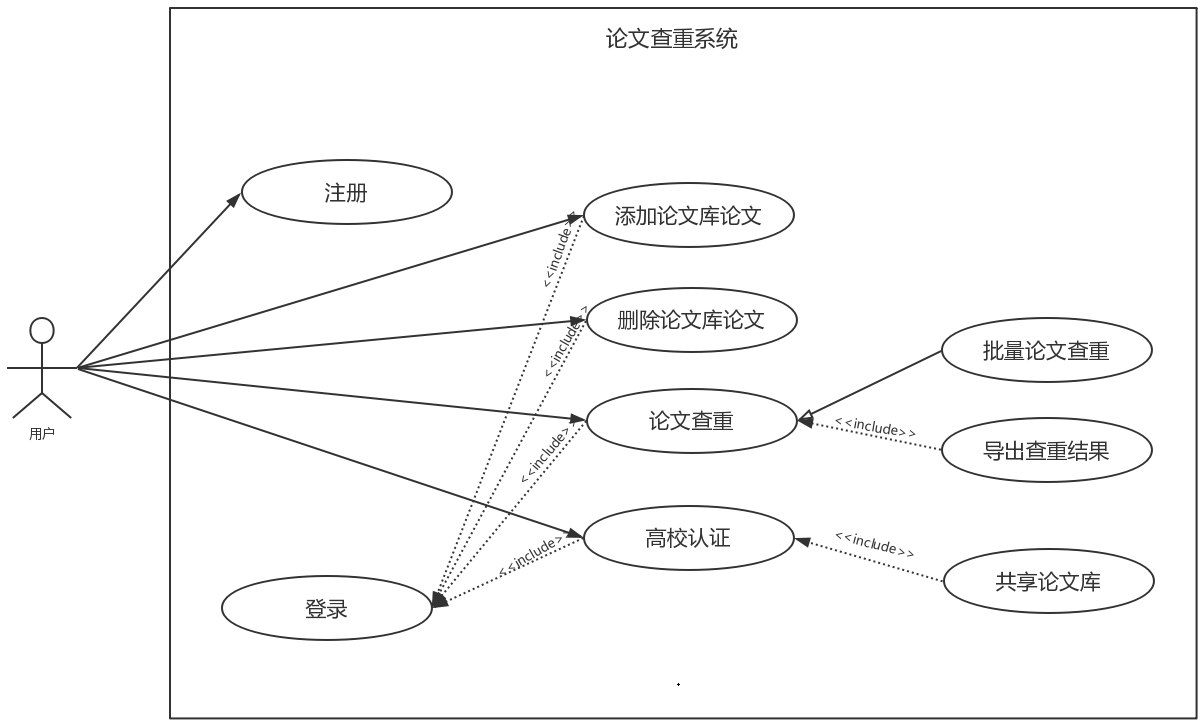


图1.论文查重系统使用用例

## 概要设计

论文查重系统主要包含以下几个模块的设计。

1、登录模块

系统架为各个高校提供服务，可供高校内部使用，但如果所有人都能使用系统，则会带来安全隐患，任何人都将可能修改论文库。故只有知道登录账户和口令的用户可以进入系统，使用系统的功能。登录模块包含账户、口令的输入界面设计和服务器逻辑层结果验证设计。

2、论文库管理模块

高校可以根据自己的需求搭建自己的内部论文库，系统允许用户自由地增加删除论文库中的论文。该模块包含论文管理界面设计、论文的添加逻辑设计、论文的删除逻辑设计。

3、查重算法模块

查重算法主要包括字符串处理算法和自然语言处理算法。前者是基于对字符串的统计进行相似度检验，后者是通过对自然语言进行建模，将文字向量化进行相似度检验。最终检验结果可以结合二者的检验结果生成。

4、文件格式转换模块

由于上传的论文因格式问题不一定能被方便地提取出内容，故在添加论文库的论文和论文查重时，要对论文进行格式转换，并把不需查重的内容剔除，保存到服务器。

5、查重报告模块

在查重算法执行完毕后，需要将统计好的查重数据进行封装，展示给用户。查重数据包括论文总重复率、相似论文及相似比例、论文重复分布图、论文主要内容的重复句子。

6、注册模块

公网系统可以提供给任何用户使用，用户需要进行注册并购买，从而能够使用系统功能。

7、高校认证模块

高等院校可以上传相关证件，经过人工审核后得到高校认证。该模块主要包括认证流程的设计、审核页面的设计。

8、高校搜索模块

用户可以搜索已经通过认证的高校，从而向高校发起论文库共享申请。该模块主要包含搜索逻辑的设计与搜索页面的设计。

9、论文库共享模块

高校间可以发起论文库共享申请，申请通过后双方可以共享论文库。论文库可以多校共享。该模块主要包括论文库合并的逻辑设计。

### 主要参考文献：

http://news.bioon.com/article/6658631.html

https://wenku.baidu.com/view/d1cf0b3943323968011c9247.html

http:// kns.cnki.net

http://www.wanfangdata.com.cn

|  |
| --- |
| 指导教师意见：  填写说明：查阅资料是否全面，提出的研究方案和计划进度是否可行，还有什么  需要注意和改进的方面，是否同意按学生提出的计划进行等。 （填写后请删除该说明） |
|  |

指导教师（审核签名）： 审核日期： 年 月