文章编号：2096-1472(2019)-05-35-04
DOI:10.19644/j.cnki.issn2096-1472.2019.05.008

面向实践教学的作业查重系统

潘理虎，张 雷，解 丹，陈立潮，赵淑芳
(太原科技大学，山西 太原 030024)

摘 要：随着高等教育从知识传授到能力培养的转型，实践教学成为大学教师普遍重视的人才培养环节。实践教学的主要难点在于教学过程的管控和形成性考核的质量。作业查重系统是降低教师负担，提高过程监控质量的有效工具。研发了一种便捷、低成本且无须数据库的作业查重系统，采用k-gram字符串匹配算法，逐字符的对比作业文档之间的字符串，当文档间相同的字符串长度达到阈值时，则计入重复度中。在软件工程课程的教学实践中的应用结果表明，教师可结合系统查重结果，对教学情况进行分析，从而实现差异化作业质量监控，提高课程质量、促进学风建设。

关键词：作业查重系统；实践教学；人机结合；学风建设
中图分类号：TP311.5 **文献标识码**：A

Research on the Assignment *Duplicate-checking* System for Practice Teaching

PAN Lihu,ZHANG Lei,XIE Dan,CHEN Lichao,ZHAO Shufang
(*Taiyuan University of Science and Technology*,*Taiyuan* 030024,*China*)

Abstract: With increasingly extensive attention of college teachers,practice teaching has become the key factor for talent cultivation,because the higher education transition from imparting knowledge to cultivating ability.The main difficulty in practice teaching lies in the management of practice teaching process and the quality of formative assessment. Assignment duplicate-checking system is effective in reducing the burden on teachers and improving the quality of process monitoring.Therefore,adopting k-gram matching algorithm to compare the character strings between assignment files,the paper proposes a convenient,low-cost and database-free assignment duplicate-checking system.When certain same character string in two files reaches the threshold value,it is counted into duplicate calculation.Practice has proven that teachers can combine the obtained results and analyze the teaching situation,thus achieving high-efficiency teaching by human-computer cooperation,which is conducive to encouraging students to innovate,improving the quality of course teaching and promoting the construction of study style.

Keywords:assignment duplicate-checking system;practice teaching;human-computer cooperation;construction of study style

1 引言(Introduction)

互联网的快速发展使得信息共享的途径增多，但电子资料的易复制性和易篡改性，也使得抄袭现象频有发生[1]。现阶段，大部分高校仍以人工审查作业为主要方式，这种方式虽然可以充分发挥教师丰富的经验优势，但对公平公正的评估学生作业仍具有很强的不确定性，且不能做到量化分析。近年来，论文查重已经成为高校中强化学术规范的重要举措，大部分学校都已购买了论文查重的新服务，但却很少有学校进行作业查重[2]。2018年9月清华发布《关于提供作业查重服务的通知》[3]，标志着清华将采用中国知网检测系统进行作业查重。考虑到使用知网检测系统进行作业查重不具有针对性，且耗费较高，于是在软件工程课程的实际教学过程中，开发了一种简便的查重系统，既可针对软件工程课程所提交的作业进行有针对的查重，又可节省教师在日常审查作业中所耗费的大量时间以提高教师的教学效率和质量。

基金项目：山西省特色优势专业建设项目2018-15，山西省教研项目J2017078，教育部产学协同项目201801128011，全国高等学校计算机教育研究会2019年度课题CERACU2019R02，太原科技大学教学改革创新项目201937.

2 作业中存在的问题(Problems in the assignment)

软件工程课程是计算机及相关专业中的一门重要课程，对于培养新工科人才具有重要意义[4]。在软件工程教学过程中，实践教学占据着重要的一部分，要充分发挥学生对该课程的主观能动性，就需要布置一些作业，以促进学生自我学习，自我思考。目前，软件工程课程的平时作业中存在数量多、难存档、难细评等问题[5]。一方面，学生们的作业大多以纸质版保存，不易长期存档，且学生们的作业存在相互抄袭现象，更有甚者，不进行思考，也不进行资料查找，直接将同学的作业进行复制粘贴，或改成自己的文件名便提交给教师。长期维持这种情况，不利于培养学生正确的学术规范，以至于在毕业的时候，习惯抄袭，最终导致难以毕业。另一方面，一个教师往往教授多门课程或多个班级，在一对多的状态下，学生提交的作业数量便成倍增长、增加了教师的工作量。当工作量很大时，教师便很难对学生的作业一一细评，尤其是难以对原创作业给予合理、公正的分数。

为解决上述问题，在软件工程实际教学中，开发了作业查重系统。该系统充分利用了现有的教学资源，可以大幅降低教学成本，在教师教学经费不充裕的情况下能够大范围的应用。

3 总体方案(Overall scheme)

软件工程课程中的作业有许多形式，如课后题、案例分析等。由于软件工程的实践性较强，大多数院校在本课程中都有项目实验的实践环节，而该实践环节的成果一般包括项目的程序代码和项目的实验报告。本系统以项目实验报告为例，对整个查重系统的流程进行介绍。具体的查重流程如图1所示，包括作业收集、作业预处理、作业查重、教师审查和结果分析。

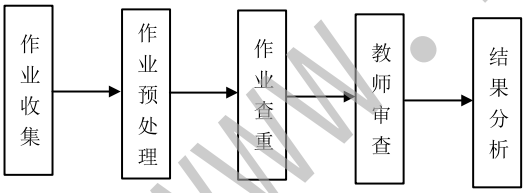

图1 查重流程图

Fig.1 The flow chart of duplicate checking

3.1 作业收集

软件工程课程实际教学中，为发挥学生的团队协作能力和培养学生的沟通交流能力，将所教授的16级计算机科学与技术专业的两个班级分组进行项目实践。由于两个班级均有41名同学，因而将每个班级的学生分为七个五人组和一个六人组。学生以小组进行实践，每个人均要提交一份实验报告。这样可以避免同组中一些能力弱的同学在项目实践过程中，不参与或很少参与项目实践，仅依赖能力强的同学进行项目开发和文档书写，但最终却能够凭借他人努力的成果通过课程考核。

传统的作业收集方式是由班长或学习委员将作业收齐，之后再统一提交给教师。这种收集作业的方式便于管理且省时省力。故本课程仍然采用这种收集作业的方式，但是为了避免因处理不同格式的文件而增加系统的复杂性，也为了便于教师对作业的审查，故要求每人的作业以统一的格式进行命名。实验报告的命名统一为“学号姓名”，且文档的后缀名统一为.docx，如“20160103张三.docx”。所有的学生将自己电子版的实验报告提交给学习委员后，学习委员初步对每个学生的作业命名格式进行检查，检查无误后将所有同学的文档放入一个文件夹中，并将文件夹命名为“班级作业名称”，如“计算机科学与技术一班实验报告”。整个文档压缩打包后发送到教师邮箱提交给教师。这样既方便教师处理作业，也便于在文档丢失的情况下进行找回。

3.2 文档预处理

教师收到提交的作业后，从邮箱下载到个人的计算机上，进行保存。之后查重系统便对实验报告文档进行预处理，也就是将文档格式处理为指定的格式，并对文档内容进行处理。预处理过程中，文档内容由python中.docx模块的Document类读取，读取的结果是分段的，并且每个段落的内容是一个document.paragraphs，之后将所有的document.paragraphs规格化。实验报告中的内容大体包含文本内容和程序代码两部分，考虑到学生的实践项目是小组共同完成的，同组学生的程序代码难免相似，且有些代码中方法名本身的字符数就很多。因此，为避免重合度虚高，实验中把所有的英文字符和标点符号过滤掉，仅保留了所要查重的主体内容——中文字符，最后将所有读取到的中文字符连接成字符串并存入.txt文档中。

按照上述方式进行预处理后的文档包括两种，一种是每个学生.docx文档所对应的.txt文档，此文档命名为“学号姓名.txt”；另一种是除每个学生本人之外的所有文档所生成的.txt文档，此文档命名为“学号姓名all.txt”。前者是为了对比单个学生之间的作业重复度，后者是为了对比学生本人和除自己之外的其他所有的学生之间的重合度。

3.3 文档查重

文档查重是整个系统的核心部分。我们称需要查重的文档为目标文档(Object Document，OD)，其中的字符串为目标字符串；与查重文档相对比的文档为模式文档(Pattern Document，PD)，其中的字符串为模式字符串。查重分为两种方式的查重，一种是学生作业之间的查重，称为一类查重；另一种是学生与除自己之外的所有文档的查重，称为二类查重。整个系统的流程图如图2。首先读取OD名，判断OD名中是否有“all”，由于只对学生本人的文档进行查重，所以OD文档中需要去除学生作业文档之外的文档。去除后，接着判断模式文档名中是否含有“all”，如果有就把模式文档名中截取的除“all”之外的字符串(简称PD1)与OD名对比，若相同，则进行一类查重；如果没有“all”则判断PD名是否等于OD名，相等则说明是学生自己的作业与自己的作业要对

比，故将查重率直接x设置为一个固定数值，为了便于最高查重率的查看，故将此数值设为0；若不等则进行二类查重。

查重流程中的一个关键点是查重算法。此系统的查重算法基于k-gram[6,7]字符串匹配算法。整个查重算法包括一个循环(Step2至Step9)，四个判断，分别称为一判断(Step4至Step9)、二判断(Step5至Step9)、三判断(Step6至Step8)，四判断(Step10至Step12)。其中，循环是从OD文档中逐个遍历单个字符；一判断是用于判断所读K-String的长度是否大于等于窗口大小(窗口大小可根据实际情况自己定义)；二判断是判断K-String是否对PD字符串进行切割，没有切割则结果等于1，此时证明该K-String是不重复的，否则的话就表明该K-String是重复的，当重复时需要对转换标志flag进行转换将其变为True；当不重复时，执行三判断，判断flag是否为真，为真则将证明除最后一位之外的K-String是重复的内容，故将其追加到重复内容的字符串中，反之，表明当前K-String的首位是不重复内容，故将其记入不重复的内容中。循环结束后，OD文档中会剩下长度不够窗口大小的未读字符串，此时执行四判断，对剩余字符串进行处理。具体算法如下：

Step1：初始化滑动窗口长度length为10，转换标志flag为False，滑动窗口字符串K-String、重复字符串Dup、不重复字符串Undup及重复内容DupText均为空字符串；

Step2：从OD文档中逐个遍历单个字符，遍历未完成时执行Step3至Step9，否则执行Step10；

Step3：将遍历到的单个字符逐个追加到K-String中；

Step4：判断K-String的长度是否大于等于length，成立执行Step5；

Step5：判断K-string字符串切割PD字符串后的结果，等于1执行Step6，否则执行Step9；

Step6：判断flag的值是否为True，为真则执行Step7，否则执行Step8；

Step7：Dup之后追加K-String字符串中的第一位至倒数第一位(不含倒数第一位)，DupText更新为原DupText后追加开始标志“<strart>”、K-String字符串中的第一位至倒数第一位及结束标志“<end>\n”，K-String更新为原K-String的最后一位和从OD中新读入的单个字符，flag转换为False；

Step8：Undup之后追加K-String的第一位，K-String更新为除原K-String首位之外的所有字符，并继续追加下一位从OD中新读入的单个字符；

Step9：flag转变为True；

Step10：判断K-string字符串切割PD字符串后的结果，等于1执行Step11，否则执行Step12；

Step11：Undup之后追加K-String；

Step12：Dup之后追加K-String，DupText在Dup追加的内容的前后分别加入标志“<strart>”和“<end>”。

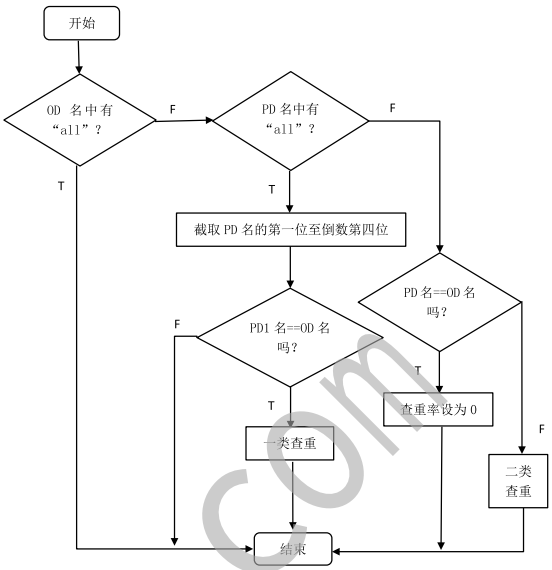

图2 系统流程图

Fig.2 System flow chart

上述步骤中，Step7和Step8是两个主要步骤，Step7之前的Step5判断出现了不一样的字符，而Step6的判断说明当前K-String是重复的，所以执行了Step7，这表明当前比较的重复内容到这里已经结束，故在Step7中开始执行下一轮的新比较。另外，Step7中的重复内容DupText内容的前后增加开头标志和结尾标志，是为了使得重复内容存入重复记录文档中时，方便教师的查看，更容易找出学生重复的内容，从而有针对性地评判每个学生的作业。Step8之前的Step5判断出现了不一样的字符，而Step6的判断说明当前K-String是不重复的，所以执行了Step8，这表明截至目前K-String的内容都不是重复的，故在Step8中将窗口后移一位。查重完成之后，将计算得到的每个学生作业文档的总字数、有效字数、重复字数，以及两类查重的重复率写入.csv格式的文档中，并将每次记录的查重详情记录到一个.txt文档中。

3.4 教师审查

教师的审查旨在查看学生作业的质量，便于进一步分析教学质量。智能查重系统最终得到的是一个.csv格式的文档和一个查重内容记录的.txt文档。其中教师可用个人计算机中的Excel软件打开.csv文档查看各个学生之间的重复率，以及学生本人与除本人之外的所有学生的重复率。

	201420010112李沛林	201420010121宋德林	201420010130曾振东	201420010132张鑫	201520010115钱子齐	201520010125颜华华	201520010132张中琦	201520010202旦兵兵	201520010210彭代伟	201520010215王怡如
201420010112李沛林	0.000	0.084	0.091	0.089	0.028	0.040	0.020	0.018	0.020	0.018
201420010121宋德林	0.067	0.000	0.246	0.394	0.140	0.017	0.014	0.014	0.014	0.017
201420010130曾振东	0.067	0.194	0.000	0.212	0.131	0.045	0.011	0.009	0.011	0.009
201420010132张鑫	0.089	0.495	0.341	0.000	0.193	0.046	0.021	0.018	0.020	0.018
201520010115钱子齐	0.057	0.349	0.442	0.391	0.000	0.035	0.046	0.038	0.056	0.042
201520010125颜华华	0.024	0.014	0.048	0.029	0.010	0.000	0.013	0.015	0.010	0.014
201520010132张中琦	0.033	0.031	0.033	0.034	0.038	0.035	0.000	0.029	0.049	0.278
201520010202旦兵兵	0.025	0.025	0.025	0.025	0.027	0.036	0.025	0.000	0.027	0.029
201520010210彭代伟	0.135	0.117	0.135	0.135	0.168	0.117	0.178	0.127	0.000	0.117
201520010215王怡如	0.015	0.018	0.015	0.015	0.018	0.018	0.149	0.017	0.015	0.000

图3 查重结果

Fig.3 The results of duplicate checking

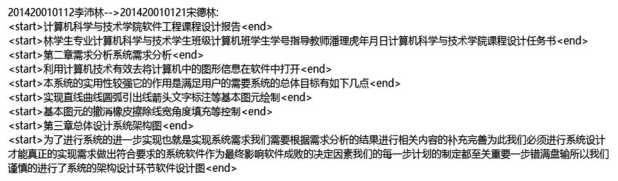

图4 查重详情

Fig.4 The details of duplicate checking

在软件工程课程中对两个班级实验报告的查重结果如图3所示。教师可用计算机中的记事本打开.txt文档，查看学生重复的内容，实验结果如图4所示。由于数据较多，图3和图4中仅截取了部分结果。通过查重结果，教师重点对那些重复率小的作业进行查看，从而筛选出高质量的原创性作业。

4 结果分析(Result analysis)

教师在进行审查后可对.csv文档中的内容进行编辑，如：生成折线图，柱状图等统计图以便查重数据的分析，也可将文档保存为后缀名为“.xlsx”的通用表格文档。如图5所示是对学生的软件工程实验报告的总字数分析的统计的分布图。可以看出，大多数学生的实验报告的字数在2000到3000，仅有四名同学字数在8000到10000。

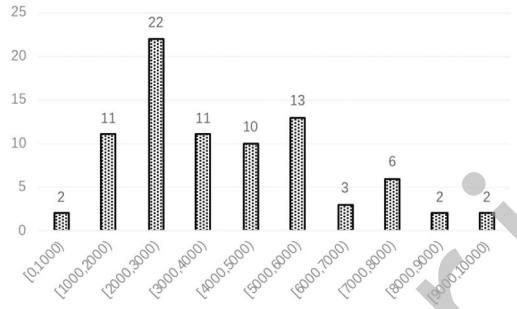

图5 实验报告总字数分布

Fig.5 Distribution of the word numbers in experiment reports

为了清晰直观地展现两个班级中实验报告的查重重合分布情况，故将查重结果分为六个等级重合度为10%、20%、30%、40%、50%及以上。且将相似度的合格线设置为30%。两类查重均需合格，否则学生作业记为不合格。

表1 查重结果统计

Tab.1 The result of duplicate-checking

重合度等级	重复率(%)	一类查重		二类查重	
		人数	占比(%)	人数	占比(%)
1	10	33	40.2	81	98.8
2	20	29	35.4	1	1.2
3	30	14	17.1	0	0
4	40	4	4.9	0	0
5	50	2	2.4	0	0
6	50以上	0	0	0	0

最终的分析结果详见表1。表中二类查重人数统计是先计算出学生本人和其余81位学生之间的81个查重率的均值，之后再按重合度等级进行统计。计算机科学与技术两个班的查重结果为表1，可以看出，大多数的学生均能自己书写文档，没有抄袭其他同学的作业。仅有6位同学的作业未能通过查重，这类学生应在平时注意引导，有近一半学生的作业查重率小于10%，可以重点进行查看，寻找原创作业。另外，对二类查重的重复率超出10%的学生的作业也再次进行查看。

5 结论(Conclusion)

根据软件工程课程实际教学情况，本文提出了面向实践教学的作业查重系统。系统利用现有的计算机进行作业的查重及分析，通过基于k-gram的字符串匹配算法得到作业的重复率。教学中，系统不仅能够提高实践教学的人力成本、时间成本、教学效果和评价质量，同时可以减少人工审查作业中的主观因素。整个系统无须数据库、无须联网，不仅最小化了软硬件开发及安全方面的投入，而且部署较为简单，易于上手，使用方便，适合大范围推广。后期可以将该系统生成可直接运行的程序，并增加可视化界面，使系统更易操作。

参考文献(References)

[1] Sidik Soleman,Astushi Fujii.Toward plagiarism detection using citation networks[C].Twelfth International Conference on Digital Information Management.IEEE,2017:12-14.

[2] Ullah Farhan,Wang Junfeng,Farhan Muhammad,et al.Plagiarism detection in students' programming assignments basedon semantics:multimedia e-learning based smart assessment methodology[J].Multimedia Tools & Applications,2018(2):1-18.

[3] 刘效仁.清华首推作业查重旨在鼓励原创[N].宁波日报,2018-09-20(010).

[4] 沙有阁,袁明磊,李晨诚.新工科背景下移动应用开发人才培养与质量保证体系研究[J].软件工程,2017,20(12):60-62;53.

[5] 陈素琴.远程开放教育课程实践性教学的研究[D].南京师范大学,2007.

[6] 殷丹平.基于CNN的代码相似度检测研究与代码查重系统[D].北京邮电大学,2018.

[7] Saul Schleimer,Daniel S.Wilkerson,Alex Aiken.Winnowing:Local Algorithms for Document Fingerprinting[C].Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data,2003:76-85.

作者简介：

潘理虎(1974-)，男，博士，副教授.研究领域：软件工程，文本处理与模式识别.

张　雷(1994-)，男，硕士生.研究领域：软件工程，文本处理与模式识别.

解　丹(1994-)，女，硕士生.研究领域：文本处理与模式识别.

陈立潮(1961-)，男，博士，教授.研究领域：软件工程，图像处理与模式识别.

赵淑芳(1978-)，女，硕士，副教授.研究领域：软件工程.