国内五个学术不端文献检测系统的对比研究

吴　凌，李海霞，郭桃美

摘　要　文章从检测技术、服务项目、所参考的数据库范围、检测结果等方面对国内常用的五个学术不端文献检测系统（中国知网科技期刊学术不端文献检测系统、万方文献相似性检测服务系统、维普论文检测系统、PaperPass、大雅相似性分析系统）进行比对和分析，探讨如何选择检测系统，并对检测系统的计算方法、数据库资源、检测结果提出思考。

关键词　学术不端；中国知网；万方数据知识服务平台；维普论文检测系统；PaperPass；大雅相似性分析系统

中图分类号　G2　　**文献标识码**　A　　**文章编号**　1674-6708（2019）235-0007-06

DOI:10.16607/j.cnki.1674-6708.2019.10.005

学术诚信是保证学术环境健康发展的原动力。然而，20世纪80年代以来，随着学术不端事件的增加和不断曝光，诸如数据造假、抄袭剽窃、代写代投等学术不端行为与国际期刊论文遭撤稿事件等频见报道，使得学术诚信受到挑战，学术诚信的重要性开始愈加受到关注。随着网络技术的发展、数据库资源的日益丰富和检索文献的便利，学术不端的表现形式也随之而发生变化，学术不端行为正在以各种新的手段不断涌现，如何去发现并定性仍然是一个难题[1]。

我国的学术不端检测软件最早于2008年研制成功并投入使用，2008—2010年经历了研制、试用到系统升级的蜕变，国内学术界对学术不端检测相关问题的关注度也在2008—2010年期间稳健上升[1]。学术不端检测系统为打击学术不端行为提供了工具和手段，智能识别技术以及科学的计算方法等为学术不端行为提供了“照妖镜”。同时，也从侧面督促了学生、学者加强原创，减少抄袭，这也为减少和避免我国学术造假作风保驾护航。相信为了避免所录用的稿件涉及学术不端行为，使用学术不端检测系统对稿件进行检测已成为编辑们必不可少的一道工作程序，这一步骤可以帮助编辑甄别抄袭文章，通过筛选来减少学术不端行为的发生。

目前国内常用的学术不端检测系统有中国知网开发的学术不端文献检测系统（以下简称知网）、万方数据知识服务平台开发的文献相似性检测服务系统（以下简称万方）、维普论文检测系统（以下简称维普）、PaperPass、超星数据库大雅相似性分析系统（以下简称大雅）等。笔者从检测技术、服务项目、所参考的数据库范围、检测结果等方面来对比和分析五个检测系统的特点，具体如下。

1　检测技术

检测技术关乎检测结果是否准确可靠。1）知网升级后的系统可以智能识别疑似文字的图片，并用OCR文字识别软件将其还原为文字进行检测；智能抓取检测文献中的公式内容进行检测，支持多公式内容检测[2]。2）万方采用的是体量最大源码和指纹库、海量数据极速源码检测引擎、高精准的源代码克隆检测算法[3]。3）维普系统采用的是动态语义跨域识别加指纹对比技术、多阶运算检测方式、云检测服务部署，使系统更快捷、稳定。4）Paperpass与大雅系统采用的均是动态指纹扫描技术。五者的核心检测技术各有特点如表1。

表1　五个检测系统的检测技术

开发商	检测技术
知网	指纹特征检测、跨语言检测、多语种检测、繁体检测、观点剽窃自动检测、表格等知识元检测
万方	体量最大源码和指纹库、海量数据极速源码检测引擎、高精准的源代码克隆检测算法
维普	动态语义跨域识别加指纹对比技术、多阶运算检测方式、云检测服务部署
Paperpass	动态指纹越级扫描技术
大雅	动态指纹扫描技术

2　服务项目

五个检测系统均有学位论文和欲发表文章的检测服务，知网、万方、维普、PaperPass均有职称论文认定的项目，知网的服务项目种类相对其他4个检测系统来说较为多样化。知网、维普与大雅均有为机构服务的板块如表2。目前，知网的学术不端检测系统不提供给个人用户使用，只开放给科技

作者简介：吴凌，编辑，广州中医药大学期刊中心，研究方向为医学期刊的编辑、校对和发展。
李海霞，郭桃美，广州中医药大学期刊中心。

期刊与社科期刊编辑部、高校研究生院部等团体用户订购使用。

表2 五个检测系统所提供的服务项目

开发商	服务项目
知网	学位论文、科技期刊、社科期刊、大学生论文、中学生作业、职称论文评审、图书专著检测、科研成果等
万方	个人文献版、硕博论文版、本科论文版、职称论文版、学术预审版、课程作业版等
维普	大学生版、研究生版、编辑部版、职称认定版
PaperPass	专科(毕业论文)、本科(毕业论文)、研究生(毕业论文)、博士（毕业论文）、职称（论文）、课程/课题（论文）、其他
大雅	本科论文、硕博论文、期刊论文、会议论文、课题内容

3 数据库范围

由表3可见，知网包含了9种数据库，万方包含了5种，维普包含了13种，PaperPass包含了5种，大雅包含了7种。每个检测系统均包含学术期刊、学位论文、会议论文数据库。维普还涵盖了高校自建资源库、古籍文献资源、个人自建资源库、年鉴资源、IPUB原创作品，是其他4个数据库所没有的。而有文献指出，大雅数据库的优势是中文图书和报纸全文数据库[4]。

4 系统支持的文献格式与检测步骤

知网的检测系统支持doc、docx、wps、caj、txt、pdf、kdh、nh格式和zip、rar 压缩包格式，万方支持doc、docx、txt格式，维普支持doc、docx、txt、pdf格式，PaperPass支持doc、docx、PDF格式，大雅支持txt、doc、pdf、docx、wps与zip、rar压缩包格式。知网和大雅均支持单篇与批量上传的方式。对于单篇检测操作步骤来说，知网系统的操作方式最简单直接，选择文件后上传即可；PaperPass系统的步骤最多，需选择文件、上传，选择文章属性、所属专业，提交共五步。

5 检测结果指标体系

由表4可见，知网、万方与大雅检测系统的计算方式相同，均是“被检测文章总重合字数在总字数中所占的比例”。从指标体系看，知网的指标体系最详细，指标维度最多，也最符合实际需求。有文献指出，高校论文管理机构最关注的指标是“去除本人文献检测结果复制比”和“相似片段分布”，其中“去除本人文献检测结果复制比”只有知网、万方、大雅的检测系统有[4]。

6 单篇论文检测结果

笔者用5个检测系统检测了同一篇文章。由表5可见，检测所得数据高到低分别是：知网20%，万方32.6%，PaperPass 38%，大雅54.4%，维普57.79%。可见知网所得的数据最低，维普得出的数据最高；万方与PaperPass、维普与大雅的检测结果接近。

7 检测结果报告

从表6可以看出，知网和维普的报告种类较多，用户可根据需求选择报告的类型。从原文对照的报告页面来看，知网、万方、维普、大雅只用一或两种颜色标注，简洁明了，方便用户一看就能辨认出相似片段内容和标注文字的性质。而PaperPass增加了段落修改功能，用户可以直接在网页上修改段落文字，即时核查出相似率，并能保存修改后的文本内容。

而高校论文管理机构关注的指标“相似片段分布”，万方、PaperPass和大雅所出具的报告中均含有相似片段分布图。

表3 五个检测系检测论文的数据库范围

数据库	知网	万方	维普	PaperPass	大雅
学术期刊	✓	✓	✓	✓	✓
学位论文	✓	✓	✓	✓	✓
会议论文	✓	✓	✓	✓	✓
报纸	✓		✓		✓
专利	✓	✓	✓		
互联网	✓	✓	✓	✓	✓
外文文献	✓		✓		✓
港澳台文献	✓		✓		
优先出版	✓		✓		
图书	✓			✓	✓
高校自建资源库			✓		
古籍文献资源			✓		
个人自建资源库			✓		
年鉴资源			✓		
IPUB原创作品			✓		

表4　五个检测系统的指标体系

开发商	指标体系
知网（12个）	检测结果(=被检测文章总重合字数在总字数中所占的比例) 跨语言检测结果 去除引用文献检测结果 去除本人文献检测结果 单篇最大文字复制比（包括单篇最大重复字数、单篇最大重复文章篇名） 重合字数(CNW) 总字数(不含参考文献字数) 参考文献字符数 大段落数(LPN) 小段落数(SPN) 前部重合度(HR) 后部重合度(ER)
万方（8个）	总相似比（=送检论文与检测范围内全部数据相似部分的字数/检测总字符数） 参考文献相似比（=送检论文与其参考文献相似部分的字数/检测总字符数） 排除参考文献相似比 可能引用本人已发表论文相似比 可能引用本人学位论文相似比 辅助排除本人已发表论文相似比（=总相似比-参考文献相似比） 辅助排除本人学位论文相似比 单篇论文最大相似比（送检文献与某一文献的相似比高于全部其他文献）
维普（6个）	全文总相似比（=复写率+他引率+自引率） 自写率 复写率 他引率 专业用语 单篇最大相似率
PaperPass（7个）	总体相似度[计算公式:（句子1相似度+句子2相似度+...+句子n相似度）/ n]句子相似度范围0.0～1.0 本地库相似度 期刊库相似度 学位库相似度 会议相似度 图书库相似度 互联网相似度
大雅（7个）	文献相似度（=送检论文中与检测范围所有文献的相似字数/送检论文正文字符数） 重复字符数 去除参考文献相似度 去除本人已发表论文相似度 文献原创度 单篇最大相似度 单篇最大重复字数

只有知网的报告中会呈现《学术期刊论文不端行为的界定标准》结果：疑似剽窃观点、疑似自我剽窃、疑似剽窃文字表述、一稿多投、过度引用、疑似整体剽窃、重复发表。并检测出是否有相似表格。

8 讨论

8.1 用户如何选择检测系统

从表5的检测所得数据可以看出，不同的检测技术、计算公式与数据库资源会影响检测结果，而且数据相差明显，面对众多的学术不端检测系统，用户该如何选择？

经过五个检测系统的对比，笔者认为，如果用户需检测用于职称评审的论文，建议优先选择知网检测系统，因目前全国人事职称评审论文管理系统应用的是知网检测系统。但目前知网的学术不端检测系统只开放给编辑部、高校团体用户订购使用，故个人用户也可选择万方、大雅，因这两个系统均有“去除本人文献检测结果复制比”和相似片段分布图，而万方检测所得数据与大雅比较，与知网更为接近。如果用户需检测欲发表文章，可选择维普或PaperPass。维普系统的数据库资源丰富，出具的报告除了有原文对照、片段对照结果外，还有格式分析报告，可以帮助作者核对是否遗漏了哪种格式的内容，并能将参考文献的格式进行自动规范，供作者复制粘贴到文档中；用户还可自行设置检测通过标准值；并可创建个人自建资源库，上传自主拥有的论文资源，在检测论文时可选择是否同时将个人自建资源库中的资源纳入检测范围，由系统自动提取解析，有效补充送检文档的比对范围。但维普检测系统得出的数据会高于其他4个系统，笔者对比了维普、知网的原文对照结果，知网报告中标注出的引用文字，维普也有标注。除此之外，维普还将疗效标准、数据、参考文献标注为复写片段，笔者推测此为得出检测数据较高的原因，用户可根据检测结果做出合理的修改。PaperPass则增加了段落修改功能，网页可见将文章内容分段落呈现，并逐句以红、橙、绿3种颜色标注，作者可以边对照结果边对段落进行修改，并能即时保存修改后的内容。

8.2 关于检测系统的思考

关于学术不端检测系统的不足之处和编辑应如何看待检测结果，许多文献均有讨论，作者不作一一赘述，只提几点思考后的疑问。

1）计算方法不相同。由表4可见，知网、万方与大雅检测结果的计算方法相同，均是“被检测文章总重合字数在总字数中所占的比例”，维普采用的计算公式是“总相似比=复写率+他引率+自引率”，PaperPass的计算公式是“（句子1相似度

表5　五个检测系统检测同一篇论文所得结果

开发商	检测结果
知网	检测结果：20%，跨语言检测结果：0，去除引用文献检测结果：16%，去除本人文献检测结果：20% 单篇最大文字复制比6.0% 重合字数(CNW)：986，总字数：4920（不含参考文献字数），参考文献字符数：730 最大段长(LPL)：164，平均段长(APL)：164 前部重合度(HR)：10.47%，后部重合度(ER)：22.43% 大段落数(LPN)：[0]-连续重合字数大于200字 小段落数(SPN)：[1]-连续重合字数大于100字，且小于200字
万方	总相似比：32.6%。检测字数：5 418，参考文献相似比：0.00%，排除参考文献相似比：32.69%，可能引用本人已发表论文相似比：0.00%，辅助排除本人已发表论文相似比：32.69%，单篇论文最大相似比：2.99%（中医治疗慢性肾小球肾炎蛋白尿的临床观察）
维普	全文总相似比：57.79%　（总相似比=复写率+他引率+自引率） 自写率：42.21%　（原创内容占全文的比重） 复写率：51.94%　（相似或疑似重复内容占全文的比重，含专业用语） 他引率：5.85%　（引用他人的部分占全文的比重，请正确标注引用） 自引率：0%　（引用自己已发表部分占全文的比重，请正确标注引用） 专业用语：0.00%　（公式定理、法律条文、行业用语等占全文的比重） 单篇最大相似度：4.11%　（原文与比对文献中单篇相似比或引用最大的比值）
PaperPass	总体相似度　38% 本地库相似度　38% 期刊库相似度　34% 学位库相似度　26% 会议相似度　15% 图书库相似度　19% 互联网相似度　7%
大雅	文献相似度54.4% 重复字符数2 338 去除参考文献相似度54.4% 去除本人已发表论文相似度54.4% 文献原创度45.6% 单篇最大相似度11.49% 单篇最大重复字数494

+句子2相似度+…+句子n相似度）/n”。因五个检测系统的计算方法不尽相同，究竟哪种计算方式更为科学与准确？笔者查阅了不少分析国内外学术不端检测系统的文献，均未提及检测系统计算方式的科学性与严谨性，因此，计算方式是否应通过专家们严格、科学的论证与多方比较，来得出科学性强和有公信力的计算方式，并拟定统一的标准？

2）数据库数量不一。作为一个学术不端检测系统，其所参考的数据库收录资源类型是否齐全、学科是否齐全、年限是否足够长、资源数量是否足够大等，会直接影响检测结果。由表3可见，数据库种类由多至少依次为维普＞知网＞大雅＞万方=PaperPass，可见五个检测系统所参考的数据库资源数量不一，相同之处为均包含了学术期刊、学位论文、会议论文数据库。根据常规逻辑，数据库资源越丰富，比对出的结果越准确。李志明[5]提出，虽然知网收录了一些字典、词典、百科全书、图录、表谱、手册、名录等，但还远远不够。另外，图片文献收录也很欠缺，如收全图书、图片等资源类型，将会对检索结果起到很好的修正作用。在收录语种方面，主要以中文为主，知网、维普、大雅检测系统均收录了外文文献资源，但与外文资源总量还有距离，为了保证科研创新性及查出相似文献，应收录足够数量的外文文献。笔者思考，每个检测系统所参考的数据库资源不一致，如何判断哪个检测系统得出的结果更准确、更科学？是否需要统一每个检测系统的数据库资源？是否统一了数据库资源与计算方法，得出的检测结果就会接近呢？因此，是否需要统一数据库种类，使不同检测系统的结果趋于接近？如何在统一数据库资源的基础上发展出各自特色，或者更具学科优势，都是值得思考的发展方向。

3）检测结果与其准确性。由检测报告比对可知，只有知网出具的报告中会呈现《学术期刊论文不端行为的界定标准》结果：疑似剽窃观点、疑似自我剽窃、疑似剽窃文字表述、一稿多投、过度引用、疑似整体剽窃、重复发表。为何其余4个检测系统的报告中没有出现类似的结果描述和字眼？笔者认为，这也从一个侧面反映出，检测系统并不能根据检测结果，准确、肯定地分辨出论文作者的引用片段是必要自引、合理他引，还是抄袭，存在对作者正常、必要的自引判定为“重复”，将引用权威著作、历史材料、法律法规等情况和统一格式的医学论文的前言、资料与方法、结果、讨论等内容判定为抄袭的情况[6]。如果盲目根据检测结果来修改文章内容，有可能破坏论文的整体内容与结构，也浪费了作者宝贵的时间和精力。

目前的检测技术对篡改、伪造、剽窃数据、伪

表6 五个检测系统的检测结果报告

开发商	报告种类	颜色标注	相似片段分布图
知网	①简洁报告 ②全文对比报告(标注出引用部分和文字复制部分，会列出疑似剽窃文字表述) ③去除本人已发表文献报告 ④全文对照报告（分段列出相似字段，并附上相似内容来源）	①红色-表示存在文字复制现象的内容 ②绿色-表示其中标明了引用的内容。	无
万方	①论文相似性检测报告简明版 ②新论文检测报告(全文比对)	①红色字体代表相似片段 ②绿色字体代表参考文献相似片段 ③蓝色字体代表可能引用本人已发表论文片段	有
维普	①片段对照报告 ②比对报告 ③原文对照报告 ④PDF报告 ⑤格式分析报告	①黑色-自写片段 ②红色-复写片段（相似或疑似重复） ③黄色-引用片段 ④绿色-引用片段(自引) ⑤蓝色-专业用语（公式定理、法律条文、行业用语等）	无
PaperPass	①综合评估 ②简明打印版 ③详细报告	①红色-代表相似度在70%以上（重度相似，请全面修改） ②橙色-代表相似度在40%～70%（轻度相似，请酌情修改） ③绿色-代表合格	有
大雅	①全文对比 ②相似片段 ③综合评估	①红色-代表重复 ②灰色-代表不参与检测 ③黑色-代表原创	有

造辅证、学术泄密等学术不端行为无能为力[7]。亦有文献指出，学术不端检测系统存在技术上的检测盲区，检测算法不够智能，无法检测翻译的外文文献，无法检测表述方式变动、语序调整、同义词替换等深层学术不端行为，只能避免“文字”抄袭而不能防止“思想”抄袭[8]。而跨语言检测技术和语义识别技术等可帮助检测系统解决“思想抄袭”问题。跨语言检测技术是通过语言规范化、候选文档检索、分类器训练、剽窃行为分析等几个步骤来进行跨语言相似性分析，该技术起步不久，在国内外均处于快速发展的阶段[9]。语义识别技术则是通过对文献的词语解析、信息抽取、时间因果、情绪判断等技术处理实现对文献的语篇理解，可很好地识别替换同义词、调整语序等学术不端行为[7]。经过比对，五个检测系统中，只有知网的检测系统得出了跨语言检测结果，在《学术期刊论文不端行为的界定标准》结果有“疑似剽窃观点”的选项。因此，检测系统应积极利用近年来蓬勃发展的新技术来壮大自己，延展检测软件的功能和使用范围，以减少误判的几率。期刊编辑可以将检测结果作为参考，但同时还是需人工介入来判定是否属抄袭或者是合理引用。

目前，有软件根据学术不端检测系统的检测结果，通过用自己的语言表述或按改变语序等行为，帮助付费者实施“一种升级的造假”[7]，使检测技术成为了一场所谓的“降重技巧”的比拼，失去了学术不端检测应有的意义。笔者在百度上输入“降重”，可以搜出相关结果约4 290万条，内容不乏降重服务、降重技巧、降重软件等，有的软件拟名“论文助手”来帮助用户降重。在淘宝网上搜索也能得出相当多的检索结果，有的商家已成功交易上万次。这种本末倒置的行为，助长了学术不端风气，严重影响了学术诚信。但正因为有市场需求，才促成了这种行为的出现和发展，要从源头来解决，并制定出相关的法律法规。

9 结论

笔者从检测技术、服务项目、所参考的数据

库范围、检测结果等方面来对国内五个常用的学术不端文献检测系统进行对比和分析，学术不端检测系统虽存在不足，但又各有特点和优势，用户可以根据自身的需求来选择检测系统。学术不端检测系统作为遏制学术不端行为的一种措施，目前发挥了其应有的作用。每位科研人员都应遵守学术规范，维护学术诚信，只有这样，研究行为才有意义，研究成果才有价值，学术环境才能健康、有序地发展。

参考文献

[1]汪雨培，王东波.学术不端文献检测技术与系统研究综述[J].江苏科技信息，2018（23）：17-21.

[2]中国知网科研诚信管理系统研究中心[EB/OL].[2018-12-03].http://check.cnki.net/vip/.

[3]万方检测[EB/OL].[2018-12-03].http://check.wanfangdata.com.cn/.

[4]孔媛媛，邓艳.知网、万方、维普和大雅论文相似性检测系统比较研究[J].产业与科技论坛，2015，14（12）：82-83.

[5]李志明.知网、万方、维普论文相似性检测系统比较研究[J].大学图书情报学刊，2015，33（1）：61-64.

[6]程翠，王静，胡敏，等.学术不端文献检测系统检测医学学术论文存在的问题及对策[J].传播与版权，2016（3）：25-26，29.

[7]郭卫兵，叶继元.学术失范、不端检测软件的功能、局限与对策——以学术研究规范为视角[EB/OL].图书馆论坛，[2018-09-30].http://kns.cnki.net/kcms/detail/44.1306.G2.20180930.1541.004.html.

[8]朱燕.试论反抄袭软件的学术规范功能及其局限性[J].兰州教育学院学报，2016（10）：91-93.

[9]刘刚，左权，杨倩茹.一种基于指纹融合的跨语言剽窃技术[J/OL].计算应用研究，2019（1）：1-10.