

论文查重的技术原理、局限及其合理应用*

康 锋 张会巍

浙江理工大学杂志社, 310018, 杭州

摘 要 为了学术期刊编辑合理应用论文查重结果开展稿件审理,阐述了查重系统的技术原理和特点,分析了其技术局限及应用局限,在此基础上提出了应用查重结果的基本原则、应用思路、审查内容和一般流程。编辑利用查重结果进行稿件审理,应遵从科研规律,尊重作者,以查重结果为线索,通过查重结果正确性核实、文字重复形式判别、写作方式判别和学术行为判别对重复内容进行审查,考察重复内容对论文质量的影响以及作者的相关学术行为,并按照特定流程对稿件作出处置。

关键词 论文查重;相似度;重复率;学术行为;抄袭

Technical principle, limitation and rational application of paper plagiarism checking//KANG Feng, ZHANG Huiwei

Abstract To assist academic journal editors to review manuscript by rational application of the results of paper plagiarism checking, this paper elaborates the technical principle and characteristics of the paper plagiarism checking system, analyzes its technical limitations and application limitations, and puts forward the basic principles, ideas, review contents and general process of application of the results of paper plagiarism checking. When reviewing manuscripts, it is necessary for editors to follow the rules of scientific research, respect the author, evaluate the duplicated contents based on the results of plagiarism checking by verification of the correctness, discrimination of different forms of text repetition, writing styles and academic behaviors, investigate the impact of the repeated contents on the quality of the paper and the author's academic behavior, and then put the manuscript into specific process.

Keywords paper duplicate checking; similarity; repetition rate; academic behavior; plagiarism

Authors' address Periodical Office of Zhejiang Sci-Tech University, 310018, Hangzhou, China

DOI:10.16811/j.cnki.1001-4314.2023.03.010

论文查重系统通过比对目标论文与文献资源库中的文本信息进行重复文字检测,可以发现目标论文是否与已有文献存在内容重复。随着大数据、云计算、自然语言处理等现代智能信息处理技术的发展,查重系统的功能日益完善,得到了广泛应用,已经成为审查论文、发现作者学术不端行为的重要工具,为学术期刊编辑审理稿件提供了非常实用的辅助手段,对于发现和遏制学术不端发挥了积极作用。然而,论文查重系统

在技术上存在一定的局限,不恰当地使用论文查重结果,会使得原本用来辅助审查论文的有力工具,异化为判断论文是否存在学术不端的特殊标准和作者规避审查的手段^[1]。

期刊编辑应用查重结果对目标论文进行审查,主要依托查重系统的各项功能开展,因此编辑只有了解查重系统的技术原理、特点及其局限,才能合理应用查重结果来审理稿件。然而,目前关于查重结果的应用研究尚不系统,没有从查重系统的技术原理和特点出发有效利用该技术,也没有提出较为系统的应用原则、思路 and 流程。因此,本文首先阐述了查重系统的技术原理和特点;在此基础上分析了查重系统的技术局限,以及编辑常用方式的局限;最后提出了编辑合理应用查重结果的基本原则、应用思路、审查内容和一般流程,为编辑合理应用论文查重结果开展稿件审查、作者学术不端行为判别和稿件处置提供参考。

1 查重系统的技术原理和特点

目前国内外已有多种论文查重系统,这些系统的技术原理相似,均采用自然语言处理技术,以文本数据为处理对象,通过文本相似度算法(下称“相似度算法”)可以快速、准确、高效地检测目标论文与已有文献的文字重复情况,为发现剽窃、一稿多投等学术不端行为提供了线索和依据。

1.1 技术原理

查重系统的技术流程可以分成2个阶段,如图1所示。第1阶段,相似度计算模型构建。首先建立文献数据库;然后对数据库中的各类数据进行预处理;最后采用特定的相似度算法,提取文本特征,建立比对特征库,从而构建相似度计算模型。第2阶段,目标论文相似度检测。对于待检测的目标论文,首先进行预处理;然后根据相似度计算模型,与比对特征库中的特征数据进行相似度计算;最后根据相似度计算结果判别重复内容,输出查重报告。其中文献数据库建立、预处理、相似度算法、查重报告输出等是查重系统的主要内容。

1)文献数据库建立。文献数据库涵盖的文献和数据资源越全面,查重结果越全面、准确。

2)预处理。相似度算法的处理对象是文本数据,

* 中国高校科技期刊研究会学术诚信与版权专项基金课题(CUJS-CX-2021-003);浙江理工大学研究生教育教学改革研究项目

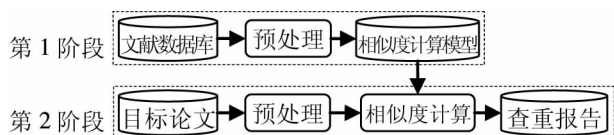


图1 查重系统的技术流程

因此查重系统在构建相似度计算模型或对目标论文进行相似度检测前,须将文献数据库中的数据和目标论文进行文本格式化、数据清理、类型标记等处理。

3) 相似度算法。相似度算法是查重系统的核心,相似度计算模型构建和目标论文相似度检测都围绕该技术展开。

文本的相似度可以从文字表达的形式和内容2方面衡量,分为表面文本相似度和语义相似度^[2]。①表面文本相似度仅以比对文本的字符匹配程度作为相似度的衡量标准。表面文本相似度算法较为简单,既不考虑词语本身的含义,也不考虑这些字词之间的关系及其在句子中真实的含义^[2],因此查重系统一般不单独使用该类算法计算文本相似度。②语义相似度主要考虑文本内在含义的相似程度,相关算法有基于字符串的方法、基于语料库的方法、基于世界知识的方法和句法分析方法等^[3]。语义相似度算法主要根据文本在语义上的某些特征或规律设计;单一算法一般只根据一种特征或规律衡量相似度,具有一定的局限,因此查重系统通常综合多种算法来计算文本相似度^[3]。

在目标论文相似度检测阶段,查重系统通过已构建的相似度计算模型,计算预处理后的目标论文与比对特征库中的特征数据之间的相似程度,得到一个量化数值,即相似度。一般只要目标论文与比对文献的相似度在句子语义级别达到一定阈值,就会被判为重复^[4]。

4) 查重报告输出。查重系统对目标论文的所有主体内容与比对特征库进行相似度计算,判别重复内容,然后统计重复情况,创建查重报告。

1.2 技术特点

1) 拥有海量的比对数据。随着文献数字化、出版数字化,越来越多的文献和科研数据被纳入了文献管理和知识服务公司的数据库中;同时,这些公司拥有较强的技术能力,可以获取海量的常用网络资源。对于国内的查重系统,文献数据库通常包含国内期刊论文、会议论文、学位论文、专利,图书、工具书、年鉴、古籍等,以及常用网站的网页资源;还包含主流外文期刊文献,查重系统可以进行跨语言检测^[4]。随着文献数据库覆盖面越来越全,查重系统的检测结果也更为全面。

2) 具有较高的检测准确度。开发查重系统的公司一般具备较强的技术能力,开发的查重系统不仅可以检测目标论文与文献数据库中字符相同的内容,还可以检测出语义相近的内容,会结合上下文进行相似度计算,对相似度大于阈值的字符复制、词语改写、语句结构调整、多篇文献组合等都可以检测出来^[4]。因此,一般查重系统可检测字符抄袭、语义抄袭等行为。

3) 记录了重复内容及相关信息。查重报告一般包含目标论文的重复字数、文字复制比(即目标论文中检出的重复字数与总文字数的比,下称“重复率”)等指标;列出了论文中所有重复内容,并可定位到目标论文和已有文献的具体位置;还记录了比对文献的出版时间以及目标论文的检测单位和检测时间,为判别目标论文与被重复文献的引证关系、作者写作行为和投稿行为提供了一定的线索。

2 查重系统的技术局限和应用局限

2.1 技术局限

查重系统受到系统设计、技术实现等因素的限制,在查准率、查全率方面存在系统缺陷,也不能准确判别重复内容在学术上的关系。

1) 查重系统通常融合多种语义相似度算法来检测目标论文是否存在重复,采用这种相似度算法尽管有利于检测文字表达形式不同但语义相同的内容,但是也会造成误检,导致查全率不是百分之百。相似度算法的阈值设置过低可能会造成误判,如表达含义完全不同的对比文字存在多个相同词语或语义相近词语时,会因相似度过高而被判为重复。图表、公式等类似具有特定文字结构和表达形式的内容,即使内容不同也会被误判为重复。预处理过程中,如果目标论文的组成部分识别不准确,可能会导致作者单位、参考文献等内容被误判为重复内容。

2) 查重系统的主要功能是检测剽窃行为。对于剽窃行为的界定,其比对的对象应当是整个人类知识。尽管查重系统开发公司掌握了海量的文献数据,但是受版权、数据获取技术和收纳时滞等因素的限制,查重系统的文献数据库中的数据,相对于整个人类知识而言毕竟是少数。查重系统对未收录的外文著作和小语种期刊等、没有及时收录的文献以及图像和音视频等非文本内容,都不能进行有效对比。另外,在相似度计算时,语义相同的改写因“相似度”不高而被判为不重复;对研究思路、方法等较为抽象的内容,查重系统也不能进行有效检测;相同内容的图表、公式等非文本数据转化为文字时,往往不能被准确识别。以上因素导致查全率不是百分之百。

3)从技术原理上看,查重系统是通过比对文本的某些特征或规律实现的,但查重结果不能自动判断目标论文中的重复内容与已有文献在学术上的关系,因此它就不可能从科学规范、作者学术动机以及内容重复造成的后果危害上作出作者学术行为的判定。这是查重系统在技术上的本质局限。

2.2 应用局限

通常,编辑应用查重结果审查稿件主要有2种方式:一种是依据目标论文的重复率等指标,判别稿件整体上是否存在剽窃;另一种是依据查重报告标记的重复内容,判别这些内容是否存在剽窃行为或表达规范问题。在此基础上,编辑根据判别结果对稿件进行处置。这2种应用方式,本质上是根据查重结果对作者的学术行为进行“界定”。因此本文分别从查重系统功能应用和学术行为界定的角度分析这2种应用方式的局限。

对于编辑依据重复率在整体上判别论文是否涉嫌剽窃,根据查重系统的技术原理和技术局限可知,有关重复率的统计是“不准”“不全”的,由其判别论文与被重复文献之间的关系也是“不准”的。尽管查重结果的重复率对审查论文有一定的提示作用^[5],但不考虑造成重复率高的原因及其对论文质量的影响,简单地通过重复率界定作者的学术行为,是草率、不合理的。将重复率作为界定作者学术不端行为的直接依据,会使作者过于在意重复率。一些作者担心论文查重出现问题,往往提前查重,故意对重复文字进行改写来“降重”,以回避编辑部的审查。同时一些学术投机分子利用查重系统的技术局限,故意抄袭小语种文献、非文本信息、不上网资源,或者对抄袭内容进行语义改写,使得查重系统失去发现学术不端行为的功能。因此,直接用查重结果对目标论文进行学术不端行为界定,会使得重复率指标异化为判断论文是否存在学术不端的特殊标准和作者规避审查的手段。

编辑依据重复内容判别论文内容是否涉嫌剽窃或存在表达规范问题,通常对于这些内容,如果是少量语句重复,编辑会要求作者修改表达方式;如果是整句、整段重复或重复内容较多时,编辑会以涉嫌剽窃为由作退稿处理。这种处理方式实质是通过目标论文中的重复内容与已有文献之间的关系来界定作者的学术行为。但是,根据查重系统的技术原理和技术局限可知,这些检出的重复内容同样存在“不准”“不全”的问题,而重复内容在学术上的关系也不能被准确判别,因此不考虑学科或研究领域的专业性和特殊性、作者的学术动机以及对论文造成的后果危害,简单依据重复内容来界定作者的学术行为往往存在问题。这种应用方

式,会促使作者进行不必要的语义改写、文献替换、过度直引甚至片面引用,造成表达扭曲,导致被引文献失去原意、引证关系不准确,甚至降低论文的论证质量。

作者学术行为的界定是一项系统工程,通常需要按照特定程序从界定对象、行为主体、行为表现和性质、界定标准等多个方面来考量^[6]。1)论文内容重复与否,与作者的学术行为是否涉及学术不端没有直接关系,依据查重报告中的重复率和重复内容来界定作者的学术行为是否涉嫌学术不端缺乏相关标准。2)在程序上,学术行为界定通常由高校学术委员会主持,由特定部门和工作人员通过调查(包括行政调查和学术评议)、谈话、处理以及接受申述、复查等特定的程序来执行^[7-8],编辑既没有相关制度作依据也没有行政职能来履行学术行为界定程序。3)由于学术论文的专业性,学术论文专业方面的判定往往需要专家介入,如果学术期刊编辑缺乏相关专业素养,则很难从专业角度对作者的学术行为进行界定。

综上所述,查重系统在技术上存在查准率不准、查全率不全的局限,且不能自动判别重复内容与已有文献在学术上的关系;在查重结果的应用上,期刊编辑缺乏相关制度、职能以及专业素养,难以按照一定程序来履行作者学术行为界定的工作。由此可知,查重结果不能直接用来界定作者的学术行为。

3 查重结果的合理应用

根据查重系统的技术原理和特点可知,查重系统是编辑审理稿件的有力工具,但在技术和应用方式上也存在不可克服的局限。本文首先提出了编辑应用该系统审查稿件的基本原则和应用思路;在此基础上,提出了依托查重结果进行稿件审查的具体内容;最后提出了审查稿件的一般流程。

3.1 基本原则和应用思路

编辑在应用查重系统审理稿件时,一方面应遵从科研规律,尊重作者,对不符合规范的内容进行检查和修正,对可能涉及学术不端的内容应履行审查、追责和更正义务,维护学术诚信;另一方面在审查稿件时,应充分利用查重系统的优势,避免其局限,合理应用查重系统的功能。

3.1.1 基本原则

编辑对论文的审理,是对论文思想、内容、方法、论证以及作者相关学术行为等多方面的综合考察,对稿件的审查和处置均建立在对论文质量和作者学术行为判别的基础之上。因此,编辑应用查重结果进行稿件审理,应遵从科研规律来考察论文在内容上是否符合

学术共同体公认的科学规范,在写作上是否符合学术表达规范,在此基础上考察作者的写作动机及其后果危害,从而对稿件作出处置。

学术论文是科研成果的表达,是一项充满主观能动性的学术创造,一篇具有洞见的论文,从文献检索、论文选题、研究设计、方案实施、结果分析到撰稿成文,每个步骤都凝聚了作者的思考和心血。草率地将论文重复率或重复内容作为判别作者行为的依据,不仅违背科研规律和科学规范,也是对作者学术工作的否定。因此编辑应尊重作者及其成果,不能想当然地怀疑作者的学术行为、随意处置论文。

学术期刊和编辑要对科研诚信承担责任^[9],在学术生态建设中发挥积极作用^[10]。国际出版伦理委员会制定的《期刊编辑行为准则和最佳实践指南》提出:编辑应当“维护学术诚信”,根据论文的重要性、原创性、研究的真实性等作出接收或退稿的决定;在处理可能的不当行为时,“对涉嫌不当行为的文章,编辑不应该只是简单地拒稿,他们有伦理义务调查受到指控的文章”,“编辑应确保对所指控的不当行为进行适度的调查”;“期刊应该有申诉机制,允许作者对编辑决定提出申诉”,“编辑应首先寻求那些涉嫌不当行为者的回应”^[11]。因此,对于稿件中可能涉及学术不端的内容,编辑有审查、追责和更正的义务,不能草率处置;在应用查重结果开展论文审查时,应当采取适当的方式,按一定流程谨慎处理。

3.1.2 应用思路

根据查重系统的原理和局限可知,编辑审稿时不能将重复率以及标记的重复内容作为判定作者行为的直接依据。正如知网学术不端文献检测系统在其系统介绍中提出的:查重系统“提供的百分比(重复率)只是描述检测文献中重合文字所占的比例大小程度,并非给定特定检测文献的抄袭严重程度”;系统的功能定位是“提供线索的工具,帮助用户更精准、便捷地发现可能存在的抄袭剽窃、一稿多投等情况,而并不给与性质上的实质判定”^[12]。显然,查重系统仅提供作者学术行为界定的线索,而不直接提供界定的“证据”。因此,在应用查重结果对论文进行审查时,应以查重报告提示的信息为线索开展相关审查。查重系统为审查目标论文与已有文献之间的关系提供了非常明确的线索,编辑可以很便捷地获取重复内容在目标论文与已有文献中的位置,从而为判别重复内容是否涉嫌剽窃或不当引用提供了辅助。

编辑对论文规范的审理主要包括2个方面:一方面,审查论文内容是否符合科学规范和学术表达规范;另一方面,根据论文内容审查结果判断作者的学术行

为,审查作者是否违反学术道德规范。因此,在根据查重报告对论文进行审理时,要考虑论文内容和作者学术行为2方面的问题。论文的内容直接体现了作者开展研究的过程以及取得的学术成果,论文的文字表达一定程度上反映了作者写作的过程和学术行为。对于查重系统检出的重复内容,应分析目标论文与被重复文献之间文字重复形式,判别目标论文与已有文献之间的引证关系,考察重复内容是否符合科学规范和学术表达规范,同时分析作者的写作行为,进而结合重复内容对论文质量的影响和作者的写作行为判别作者的学术行为,最终根据作者的学术行为对稿件进行处置。

3.2 审查内容

根据查重技术的特点、局限以及查重结果的应用思路,编辑在开展论文审查时,主要的审查内容包括查重结果正确性核实、文字重复形式判别、作者写作方式判别和作者学术行为判别。

3.2.1 查重结果正确性核实

编辑应用查重结果时主要针对标记的重复内容,但这些内容可能存在误检,因此首先应对查重报告提示的信息进行正确性核实。主要包括:1)排除查重系统在预处理阶段由于目标论文结构内容识别错误而造成的误检。2)对目标论文存在的重复内容与被重复文献进行对比,排除因相似度算法的局限而被误识别的部分。排除误检内容后保留确实存在文字重复的内容,为后续分析做好准备。

3.2.2 文字重复形式判别

正确性核实后的文字重复内容,主要有以下几种重复形式。

1)字符重复。字符重复是指目标论文与被重复文献的相似内容,在文字及其结构上完全一致。这种情况也称为文字复制。字符重复达到句子层级以上的内容,一般在语义上也重复。

2)语义重复。语义重复是指目标论文与被重复文献的相应内容,虽然文字不完全一致,但其表达的含义相同或相近。

根据查重系统的原理,查重系统检测出的文字重复只有字符重复和语义重复2种形式。

3)隐性重复。隐性重复是指目标论文与被重复文献的相应内容,虽然查重系统由于技术局限没有检测出来,但是实际上语义相同。显然,查重报告不能显示存在隐性重复的内容,但其属于查重系统的应用目标,也是编辑需要关注和防范的内容,所以本文也对此进行分析。

3.2.3 作者写作方式判别

1)自我表达。自我表达是指作者用自己的语言

表达自己的观点,即在写作动机上作者表达的内容与已有文献无直接关系。一般来说,作者表达的句子层级以上的语句与已有文献存在字符重复的可能性很小;但正确性核实后的重复内容,也有可能存在语义重复,从而被查重系统检测出来。

从作者主观动机上看,这种写作方式不涉及剽窃。出现论文与已有文献语义重复,主要有3种情况:①论文与已有文献在表达内容上相似,但这些内容属于同一研究主题,如引用了类似文献,应根据引用文献和文字重复形式进行引用方面的审查;如属于专业常识,可以不标引、著录文献。②论文与已有文献确实存在引证关系,但作者没有标引必要的文献,“该引而未引”,本文将此归为引用缺失这种引用不当的情况。③论文的核心观点、结论与已有文献存在语义重复,可能是目标论文缺乏创新性,因此需要编辑联合专家进行论文的原创性、先进性方面的审查。

2)引用。作者写作时参考了已有文献,主要存在合理引用、引用不当2种情况:①合理引用是指目标论文引用了已有文献,且文字表达、文献标引和录著都符合规范;如果引用众所周知的且不具备特定表达形式的专业常识,可不标引、著录文献。②引用不当主要包括引用形式不规范、引用缺失、过度引用、片面引用、转引等。引用形式不规范是指论文中文献标引或著录不规范;引用缺失是指未标引、著录引用文献,该引而未引;过度引用是指不必要的直接引用或引用过多文献,从而影响论文的原创性或论证质量;片面引用是指改变被引内容的本意,故意歪曲作者原意;转引是指将引自其他文献的引文标注为直引。

3)抄袭。从查重系统检出的重复内容看,文字重复主要包括字符重复和语义重复2种形式,对应的抄袭形式是字符抄袭和语义抄袭。字符抄袭是指作者抄袭的内容与已有文献的文字及其结构均相同;语义抄袭是指作者抄袭的内容与已有文献的文字不完全一致,但其表达的含义相同或相近。

4)隐性抄袭。查重系统受技术局限,未能检测出目标论文与已有文献存在的重复内容,即存在隐性重复。作者出于投机心理,通过抄袭未上网的文献,或通过语义改写、多重翻译等方式对已有文献进行处理,或通过事先查重然后根据查重结果将重复文字进行改写,使得查重系统不能准确检出重复内容。

3.2.4 作者学术行为判别

编辑依据查重结果对作者学术行为的审查,实质上是根据目标论文与已有文献之间的关系和表达方式,来界定作者的学术行为。有关学术不端及相关概念的定义较多,本文参考美国科技政策办公室认定科

研不端的标准^[13]、《澳大利亚负责任研究行为准则》^[14]与和鸿鹏等^[15]对科研不端认定的研究,从科学规范、主观动机和危害后果3方面的因素来判别作者的学术行为。针对学术论文内容重复涉及的科学规范、作者学术行为及危害后果,本文将作者学术行为归为学术不当、学术失当和学术正当3类;具体类型主要根据《学术出版规范期刊学术不端行为界定》^[16]进行判别。

1)学术不当行为。学术不当行为是指作者在论文写作和从事相关科研过程中,为促成论文发表而故意违反学术共同体共同认可和遵守的学术道德规范和学术表达规范,且对论文的学术质量及其出版造成严重影响的行为。在作者应用已有文献开展写作的行为中,涉及学术不当行为的写作行为主要是抄袭和隐性抄袭。

①无论是字符抄袭还是语义抄袭,都是作者故意将他人的观点、数据、图像、研究方法、文字表述等挪为己用。具体类型包括:整体剽窃、重复发表、未发表成果剽窃等学术不当行为,以及观点剽窃、研究方法剽窃、数据音视频剽窃、文字表述剽窃等行为^[16]。

②隐性抄袭这种写作行为,带有明显的主观恶意,可能涉及整体剽窃、重复发表、伪造、篡改等,属于学术不当行为^[16]。

从后果危害来看,无论是抄袭行为还是隐性抄袭行为,本质上都是作者故意将他人观点、数据等挪为己用,如果对论文的真实性、原创性、论证性等及其出版产生较大影响,甚至损害原作者的署名权、首发权等著作权,应当判别为学术不当行为。

2)学术失当行为。学术失当行为是指作者由于缺乏必要的知识而造成失误、疏忽或无意识地违反科学规范的行为,且对论文的学术质量及其出版存在负面影响,但经过一定努力可以修正。在作者应用已有文献开展写作的行为中,涉及学术失当行为的主要是引用不当行为和部分抄袭行为。

①由于引用不当影响论文的规范性、原创性或论证质量,应判别为学术失当行为;如对论文的原创性产生较大影响,编辑宜联合专家进行论文的原创性、先进性方面的审查。

②对于抄袭行为,如果只涉及论文的部分论点、数据或文字表述等,对论文的原创性、论证质量、规范性影响不大,且经过一定修改可以解决的,宜判别为学术失当行为,编辑应提示作者改正并规范引用。

3)学术正当行为。学术正当行为是指作者的学术行为符合学术道德规范,论文写作符合学术表达规范的行为。显然,在作者应用已有文献开展写作的行

为中,合理引用是唯一的学术正当行为。

综上,应用查重结果对论文审查的主要内容如图2所示。

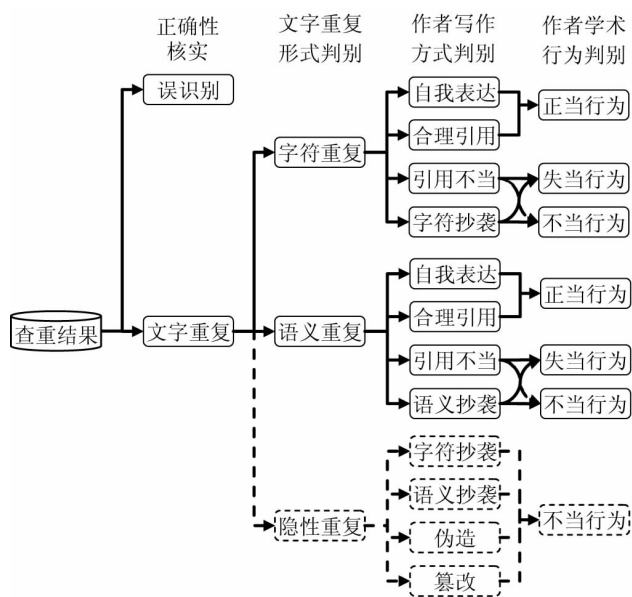


图2 应用查重结果开展稿件审查的主要内容

3.3 一般流程

编辑对稿件的审理是一项系统工程;编辑依据查重结果对论文的审理,需要考虑各方面的因素,采用合理和严谨的流程,针对查重系统检出的问题,在查重信息正确性核实、重复形式判别、作者写作方式判别和作者学术行为判别的基础上,考虑后续处置措施。应用查重结果审查稿件的一般流程主要包括编辑审查、稿件处置和作者申诉3个阶段,如图3所示。

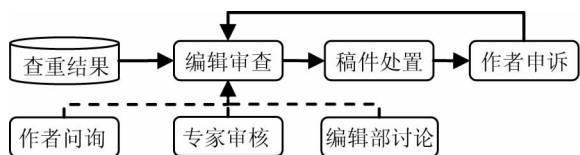


图3 应用查重结果审查稿件的一般流程

1) 编辑审查。编辑对目标论文进行查重,获取查重结果,然后依据查重结果对论文进行审查,主要审查内容见图2。在编辑审理稿件和判别作者学术行为时,对于涉嫌学术不当或学术失当的内容,可参考《期刊编辑行为准则和最佳实践指南》^[11]、《科研诚信新加坡声明》^[17]、《科研诚信案件调查处理规则(试行)》^[7]等按一定流程开展审查。

如果作者涉嫌学术不当或学术失当行为,且编辑根据查重报告等途径获取的相关佐证材料难以确认作者学术行为的,宜通过作者问询、专家审核、编辑部讨论等方式开展审查:①在涉及观点和方法剽窃上,应给

作者申诉的渠道,在听取作者解释后再决定后续处置措施。②如果涉及专业性问题是编辑不能确认的,那么可向作者问询,接收作者答复、整理相关佐证材料,由专家进行审核,再根据专家意见作出处置。③对于某些特殊问题,如果编辑不能依据相关标准和编辑部规定判定,宜提交编辑部或编委会讨论,再作出稿件后续处置决定。

2) 稿件处置。编辑对稿件的处置,需考虑作者的学术行为以及作者修正难度、编辑审理难度、论文发表时滞等多个因素。具体的处置措施,依据稿件审理所处的阶段,可分成稿件审理阶段和已发表阶段。

如稿件在审理阶段,可以参考以下方式处置:①对于涉及学术不当行为且影响恶劣的,退稿,编辑部备案,并通报作者单位或相关行政管理部门。②对于涉及学术不当行为且影响较大的,退稿并警告作者,编辑部备案。③对于涉及学术失当行为的,警告或提示作者,并请作者修正;如果反馈给作者后,作者不能在短时间内解决而影响出版的,那么这类稿件宜先退稿,请作者解决问题后重新投稿再处理。

如稿件已发表,可以参考以下方式处置:①对于涉及学术不当行为且影响恶劣的,撤稿并警告作者,编辑部备案,并通报作者单位或相关行政管理部门。②对于涉及学术不当行为且影响较大的,撤稿并警告作者,编辑部备案、通报。③对于涉及学术失当行为且论文修改内容较少的,请作者修改后对论文进行勘误处理;如果稿件修改内容较多的,请作者修改后对论文进行替换。

3) 作者申诉。对于作出退稿、警告、通报、备案等处置决定的,应当向作者反馈处置决定,同时接受作者申诉。如果作者提出申诉,那么应结合作者申诉内容和佐证材料重新审查相关内容,再根据审查结果作出处置。

4 结束语

学术期刊编辑是学术论文的把关人,作为学术共同体的一员,应做好学术出版的质量和规范工作,努力使学术诚信建设落到实处。

期刊编辑部、高校及有关教育主管部门越来越广泛地使用论文查重系统,使得众多学术期刊、作者和有关教育主管部门产生了一定依赖。在各类网络平台的助推下,一些作者过于在意查重结果,甚至利用相关漏洞规避学术不端检查,导致了更为严重的问题。关于这一问题的治理有待后续研究。

5 参考文献

- [1] 马亮. 学位论文查重的“七宗罪”[EB/OL]. (2020-04-

- 08) [2020-04-08]. <https://www.163.com/dy/article/F9N1J7460534A4SC.html>
- [2] 王春柳, 杨永辉, 邓霏, 等. 文本相似度计算方法研究综述[J]. 情报科学, 2019, 37(3): 162
- [3] 陈二静, 姜恩波. 文本相似度计算方法研究综述[J]. 数据分析与知识发现, 2017, 1(6): 5
- [4] 《中国学术期刊(光盘版)》电子杂志社有限公司. 常见问题解答[EB/OL]. [2022-11-06]. <https://check.cnki.net/Article/about/2009/03/25.html>
- [5] 郑晓梅, 张利田, 王育花, 等. 期刊编辑和科研人员对学术不端及其边缘行为的界定、防范和处理认知的调查结果分析[J]. 中国科技期刊研究, 2020, 31(4): 402
- [6] 李丹丹, 赵永峰. 研究生学位论文检测与查处: 问题与改进[J]. 重庆高教研究, 2020, 8(5): 106
- [7] 科技部等二十部门关于印发《科研诚信案件调查处理规则(试行)》的通知[A/OL]. [2023-03-30]. http://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1779/202007/t20200715_472861.html
- [8] 中华人民共和国教育部. 高等学校预防与处理学术不端行为办法[A/OL]. [2023-03-30]. http://www.moe.gov.cn/jyb_xxgk/xxgk/zhengce/guizhang/202112/t20211206_585094.html
- [9] 缪弈洲, 张月红. 学术期刊对科研诚信要承担的责任[J]. 编辑学报, 2021, 33(4): 468
- [10] 张晓雪. 科技期刊在学术生态建设中如何发挥积极作用[J]. 编辑学报, 2021, 33(1): 23
- [11] 国际出版伦理委员会. 期刊编辑行为准则和最佳实践指南[R/OL]. 苗晨霞, 译. (2012-05-16) [2022-11-08]. <http://www.cujs.org.cn/site/content/6105.html>
- [12] 《中国学术期刊(光盘版)》电子杂志社有限公司. 科技期刊学术不端文献检测系统(AMLC)[EB/OL]. [2022-11-08]. <https://check.cnki.net/amlc/>
- [13] Federal research misconduct policy[EB/OL]. (2000-12-06) [2022-11-08]. <https://ori.hhs.gov/federal-research-misconduct-policy>
- [14] Australian Research Council. Guide to managing and investigating potential breaches of the Australian code for the responsible conduct of research[EB/OL]. (2018-06-14) [2021-10-21]. <https://www.nhmrc.gov.au/about-us/publications/australian-code-responsible-conduct-research-2007>
- [15] 和鸿鹏, 齐昆鹏, 王聪. 科研不端认定的依据与争议[J]. 北京航空航天大学学报(社会科学版), 2022, 35(1): 77
- [16] 学术出版规范 期刊学术不端行为界定: CY/T 174—2019[S]. 北京: 中国书籍出版社, 2019
- [17] 科研诚信新加坡声明[R/OL]. 《中国科学》杂志社, 译. (2010-09-22) [2022-11-08]. <https://www.wcrif.org/documents-singapore-statements/translations-statements/26-singapore-statement-chinese-csb/file>

(2022-11-30收稿;2023-03-31修回)

用于计量和计数的个位数字宜使用阿拉伯数字吗?

GB/T 7713.2—2022《学术论文编写规则》规定:“鉴于阿拉伯数字具有笔画简单、结构科学、形象清晰、组数简短、国际通用等优点及科技语言的特殊性,论文中数字使用的总原则是:凡是可以使用阿拉伯数字,而且又很简明清晰的地方,宜使用阿拉伯数字。”对于“计量和计数的数字”,“为达到醒目、易于辨识的效果”,强调“应使用阿拉伯数字”,并列举了“应写作”“4人”的示例。这就明确地告诉我们,对学术论文中用于计量和计数的数字(包括个位数字),国家标准是建议采用阿拉伯数字的。

GB/T 7713.2—2022对数字用法的标准化,为我们做出了示范,标准中凡是可以使用阿拉伯数字的计量和计数的数字,都使用了阿拉伯数字。为便于编辑同人学习参照,我们特意将使用了阿拉伯数字的个位数字的案例摘录如下。

“本文件是GB/T 7713的第2部分”;

“1987年版第3章、第4章”;

“论文一般包括以下3个组成部分”;

“每篇论文以选取3~8个关键词为宜”;

“正文部分应根据需要划分章节,一般不宜超过4级”;

“各层次章节编号全部顶格排,其后空1个汉字的间隙接排标题”;

“只有1幅插图、1个表格时,应编为‘图1’‘表1’”;

“描述传递现象的特征数由2个字母组成”;

“8个二进制倍数词头符号应分别为……”;

“如长度0~20 m平均分为4组”;

“大于999的整数和多于3位数的小数”;

“2个数字连用表示的概数”;

“在数学式中宜避免使用多于2行的表示形式”;

“引题小5号黑体,内容小5号仿宋”“编号、标题小4号黑体,内容5号宋体”。

(郝 远)