遏止学术不端行为　保护科研原创成果*
——《浙江大学学报(英文版)》作为 CrossCheck 中国第一家会员的实践与体会

林汉枫　张欣欣　翟自洋　伍秀芳　张月红**

收稿日期:2008-04-07
修回日期:2009-06-15

《浙江大学学报(英文版)》编辑部, 310027　杭州市浙大路 38 号, E-mail: jzus@ zju. edu. cn

摘　要　文章介绍了在线检测英文文献原创性的反剽窃工具 CrossCheck,具体分析和评述了《浙江大学学报(英文版)》使用 CrossCheck 检测时所遇到种种不合理的学术现象,如重复发表、自我抄袭、搬来主义、东抄西凑和随意摘用等,同时探讨了不同论文体裁和时滞效应问题。

关键词　CrossCheck　原创性　版权　学术不端　在线检测

1　引言

历史的进程已经见证科技作为第一生产力加速了社会的发展;历史的文明同样也在叙说着一个事实,即创造精神财富和推进科学进步的知识群体也是社会的一个组成部分,他们中有些人或许能突破科学研究的禁区,但未必能抵御与生俱来的各种利益诱惑,严守学术道德的底线。2008 年,英国《自然》的一项调查报道指出,每年约有 3% 的研究人员被发现有科学上的不端行为,主要表现为伪造数据及剽窃,尤其是医学领域的论文剽窃尤为突出[1]。再如,Mounir Errami 和 Harold Garner 对美国医学索引(Medline)数据库中 1975 年至 2005 年大约七百万篇生物医学论文摘要用 eTBLAST 进行分析,发现平均 7 万多篇出现高度的相似,有重复发表之嫌,且该现象呈现逐年严重的趋势;他们还发现论文重复率与各个国家发表的论文数大致呈正比关系,在占据其数据库 75% 的主要来自美国、日本、德国、中国、英国、意大利、法国和加拿大等国家的论文中,重复现象令人担忧,尤其中国和日本,其论文重复率是其发表论文数所占比例的两倍多,追溯其原因或许是不同文种的译制本,以及文化道德标准的差异所致[2]? 学术剽窃的行为除了表现为抄袭他人的研究成果之外,自我抄袭的现象也不容忽视。Sorokina 等人通过对 arXiv 预印本网站上 284834 的文献进行检测分析,发现可疑的剽窃和自我抄袭分别占0.2% 和 10.5%[3]。

面对这些学术不端行为,除了需要主观上不断陶冶情操和提高学术修养,客观上应该实施遏制举措。不少学术期刊已经使用强硬手段规范作者的行为,例如,规定在投稿的一定期限内不得一稿多投,或者要求签署版权委托书等有法律效应的合同,一些国际学术出版商还采取了不同的措施去发现并防止这些学术不端行为。例如,大多数的学术出版商采用手动搜索或者依靠专业领域内审稿人的经验去判断所提交的文章有否剽窃行为;英国医学出版集团(BMJ)的一些资深编辑通过 Google 快讯辅助检查剽窃结果;爱思唯尔(Elsevier)出版公司发布了出版道德资源手册(PERK)为期刊编辑提供即时和广泛的在线支持[4]。然而,这些方法必须建立在审稿人或编辑对文章内容产生怀疑的前提下,知道从何处下手。那么,是否有更有效的方法能达到有力地反剽窃的目的呢?

CrossCheck 就是在这样的背景下出来的一个基于网络平台的自动检测文档原创性的反剽窃工具。它是由 CrossRef 和 iParadigms 联手在 iThenticate 技术基础上研制的,服务于

* 国家自然科学基金重点学术期刊资助项目(No. 30824802)和中国高校科技期刊学研究项目(No. GBJXC0819)
** 通讯作者: 张月红

CrossRef 的成员。在国际出版链接协会(The publishers International Linking Association-PILA)牵头下,国际几大出版商和电子电气工程师协会(IEEE)及美国计算机学会(ACM)共同参与了这项全球性项目。正是由于 CrossCheck 能够在全球范围内最大程度地检查和防范学术剽窃行为,达到严正学术道德,净化学术空气的目的,使其一举赢得了全球学术与专业出版者协会(ALPSP)颁发的 2008 年度全球最佳出版创新奖[5]。目前全球会员有 50 多家,包括一些国际科学出版集团和科学学会:自然出版集团(NPG),爱思唯尔(Elsevier),施普林格(Springer),威立·布莱克威尔(Wiley-Blackwell),英国医学期刊出版集团(BMJ),泰勒弗朗西斯出版集团(Taylor & Francis),美国科学进步协会(AAAS),美国物理学会(APS)等[6]。《浙江大学学报(英文版)》在国家自然科学基金的重点期刊项目的资助下,于 2008 年 10 月正式成为中国第一家 CrossCheck 会员。在继续坚持全面严格的国际审稿体制的基础上,把 CrossCheck 作为学术把关的"第三只眼睛",与全球的科学家和出版者们共同合作,为学术期刊质量和尊严尽职尽责。

下面简单介绍 CrossCheck 的工作原理,并结合工作中所遇到的各种不合理引用和抄袭现象进行归纳分析,提出一些问题与同行商榷。

2 CrossCheck 功能与特点

2.1 工作原理

CrossCheck 的工作原理由两个部分组成:一是基于全球学术出版物所组成的庞大文献资源,包括其储备的数据库和互联网开放共享资源[6];二是基于网页的在线检验系统,将上传的英文文本与文献资源作对比,自动在线产生被检测文本与各匹配文献相似度的原创性报告。工作人员可据此报告结果分析判断文件的原创性和引用的合理性,进而发现和指出是否有抄袭或剽窃和重复发表等各种不端行为,维护原创作者的著作权。

2.2 操作介绍

CrossCheck 的操作简单易行。其界面如图 1 所示,从左到右依次为文件名(Title),相似度报告(Report),作者名(Author),操作日期(Processed)和操作选项(Action)。其中的相似度报告数据为该被检测文献对应的所有对比匹配文献的相似度的总和,当其总量超过 50% 时,系统会自动显示黄色背景,提醒操作者的注意。只要点击其中的相似度数据,系统便直接进入具体报告列表,在此操作者可以对论文具体的"文本重叠"现象进行分析判断。其中,界面的左栏为上传的被检测文本,凡与之匹配的对比文献相似部分系统以相同的颜色和序列号标识;右栏的每个单篇匹配文献的相似度大小顺序排列。界面的上方指出了总相似度(Similarity Index)和视图选项(View),含有四种选项:相似度报告(similarity report)、内容追踪(content tracking)、概要报告(summary report)、和最大匹配(largest matches)。点击左栏相似之处,右栏会自动转换成与其匹配的对比文献全文(不包括图表),从而对比两文件之间的具体相似情况;点击右栏对比文献处,则进入该文献所在的页面。如果它属于开放存取的期刊或者使用者购买了的数据库,则可以查看全文,进一步对其中的图表公式等具体内容的相似性进行分析。

2.3 优势与不足

作为在线检测工具,CrossCheck 不需要下载安装程序,因此不受电脑系统和时间地点的限制;可检测的英文文件类型多样,包括 Word 和 Word XML, WordPerfect, RTF, HTML, Text, PostScript 以及 PDF 文档[7];检测速度快,根据文件的篇幅(50 页以内)不同,在几十秒至几分钟之内便可产生报告结果;最重要的是,CrossCheck 可以对文本全文进行比较全面的对比检查,同时可供对比的资源极其丰富。与其他反剽窃工具相比,CrossCheck 有着不可取代的优势,如 iThenticate 虽然是 CrossCheck 的技术基础,但是它只能对文本的前后部分内容进行检测[6],不够全面;又如 eTBLAST 检测对象局限于美国医学索引数据库中的摘要[4]。

作为新生事物,CrossCheck 不可避免地存在一些不足之处。例如,CrossCheck 只能对论文的文本进行检测,而其中的图片、表格和数学公式等则需要操作者进一步查找原文进行详细分析;CrossCheck 是对语言表达的原创性进行分析,但它并不能直接发现内容的造假,需要结合其他的辅助手段进一步分析判断;目前,CrossCheck 检测的文件语言只局限于英文。可喜的是中国学术期刊(光盘版)电子杂志社和同方知网有限公司联合开发的"科技期刊学术不端文献检测系统"(AMLC),可用于检测中文文献的不端学术现象,弥补其部分不足[8]。

图 1　CrossCheck 在线系统检查文档产生的原创性报告界面

3　实例分析与讨论

《浙江大学学报(英文版)》自从成为 CrossCheck 的会员后,一直坚持在国际同行评审和正式发表之前对论文先后进行至少两次 CrossCheck 的检查。我们从以下几个问题的不同学术角度进行分析和评注,以求同行共识。

3.1　五种学术不端现象

在几个月的工作实践中,我们通过 CrossCheck 对不同论文体裁相似度标准的初步确定和分析,发现大多数的论文作者是秉着严谨治学的态度,其论文相似度比较低。然而,约有 20% 的文章由于各种各样的原因具有一定程度不合理的摘用他人和自我抄袭等现象,其中约 5% 的文章甚至涉及剽窃和侵犯版权之嫌。根据不同现象的表现形式,我们归纳出以下五种不合理现象。

(1) 重复发表

在 CrossCheck 检查中,发现一些作者为了增加文章发表的几率而一稿多投,又或为了增加发表的文章数,将已经有正式书刊号的会议论文集或者电子期刊上发表过的文章,经过略微增减修饰,或原文不变地重新向有关期刊投稿。这些行为违背了期刊的发刊原则,损害了多个期刊的利益,造成了出版资源的浪费。如,有一篇来自布基纳法索和法国某作者的文章经过 CrossCheck 检查,虽然没有在正文中发现明显的相似度,但是参考文献完全一致,因而与其中单篇匹配文章的相似度高为 18%。异常的现象促使我们进一步对其全文的图表数据进行详细分析对比,发现该作者一年前曾在某刊上发表过类似的文章,两篇论文的重复性高达 80%,包括一模一样的三个图和一个表格数据,可见完全是旧数据、老资料的重新发表。另外,还有一个代表性例子,我们在 CrossCheck 平台上对某一即将发表的文章进行最后的检查时,发现完全相同的摘要,进而得知该文作者的博士论文专集已经在线某国大学出版社的电子资源库,而且与其相关的核心部分已经在五年前发表两篇论文。虽然作者认为现在的投稿是博士论文专集的一部分仍可以发表,但是我们认为学科具有即时反应的特点,经过五年的时滞后,论文没有补充新的进展,没有创新,也就失去了重新发表的意义。当然相关的界定标准还有待讨论。

(2) 自我抄袭

CrossCheck 检查中发现的另外一种比较典型的现象就是论文作者的自我抄袭或者相互抄袭,不可避免地导致高相似度产生。相当一部分作者持有这样一个观点,即已经在其他刊物发表过的内容,基于同一个研究课题的不同方向,由于采用的材料设备和方法相同,因此在新投稿中,尤其是引

言、材料方法和讨论部分重复使用已经发表的内容应该是合理的。这种论调从理论上是经不起推敲的,因为已经发表的文章意味着原创内容已经公诸于众,再次充填于新文中对作者本人有拼凑之嫌,对期刊来讲是浪费出版资源,更重要的是对读者来讲,严重地浪费了他们的时间去面对这些有失原创水准的章节。如果重复用一些相似的内容写一篇新的论文,何不把两篇合并成一篇高质量的文章发表呢? 爱思唯尔科技部中国区副总裁安诺杰指出:“研究人员会发表一系列相似的论文,这对期刊编辑来说是件很痛苦的事情。因为这五篇文章实际上可以写成一篇很好的文章。在美国或者欧洲却很少出现这样的情况,因为当科学官员或者基金官员在查看出版记录时,他们会看具体的论文,如果发现五篇论文都有相似的主题和重复的内容,他们会认为这样做是不严肃的”[9]。**科学的重要性不是通过数量来衡量的,而是要看论文本身的质量。**

(3) 搬来主义

还有一种现象在生物医学领域的论文中比较常见,如作者直接复制他人的试验方法和操作描述,然后在试验条件和数据上更改替换。有些作者坚称这种试验方法借用经典文献的表达是正常的,且在多个期刊上常见。对此,我们与国际同行进行了讨论,并研究了一些国际知名期刊(如《科学》(*Science*)、《自然》(*Nature*)等),几乎没有发现类似现象。而且在理论上,这种说法也是不合理的。虽然大多数的科研是在结合前人研究成果的基础上,借鉴或重复他人成功的方法试验新的材料,探讨新的结果,但是作者在撰写论文的时候,除了应该引用经典文献外,应该根据时间、地点和条件,用自己的语言去描述自己的工作,总结自己的观点,因为语言描述也是一种原始创造过程。

(4) 东抄西凑

在 CrossCheck 检查中发现有极少数文章,几乎全文大部分的内容都能找到与之相匹配的对比文献,甚至大段内容的相似,甚少用自己的语言描述。造成这种情况很可能是因为作者整篇文章的写作是靠东抄西凑完成的。这样的论文反映了作者对待科学研究极端不严谨的学术态度。

(5) 随意摘用

此外,在 CrossCheck 检查中还发现有些作者在引述他人的观点或者描述他人的科研成果的时候,文章中大段的句子与匹配的对比文件相似,却没有文献出处。这样会误导读者认为是作者自己的观点,极易引起版权纠纷。这种取向或许是因为作者英文写作上的先天不足而去模仿他人的表述。他们从主观上并没有抄袭之念,只是在写作的时候没有做到模仿量的恰如其分。还有一种现象是一些作者认为只要注明了文献出处,就可以直接大块照抄他人之段落,这些表现在认识上有误区[10]。哈佛大学关于“抄袭”的规定指出:“如果你的句子与原始资料在观点和句子结构上都非常相似,并且结论与引语相近而非用自己的话重述,即使你注明出处,这也是抄袭。你不能简单地改变原始资料中的几个词语或者对其进行摘要性重组,你必须用自己的语言和句子结构彻底地重塑你的总结,要不就直接引用。当然对于已经成为学术界的常识经典名句、即使不做说明也不会对提出者的归属产生误会的观点,则可以不注明出处[11]”。

3.2 不同体裁的相似度问题

值得强调的是,论文体裁不同,其相似程度也有所不同。如研究论文,尤其是以科学创新和独特见解为特点的科学快报,作者在撰写文章的时候应该充分体现出文章的原创性;而综述性论文是对某学科历史和现在的科研成果进行总结评述,并结合自己的观点进行分析和讨论,引用参考文献较多,因此,论文总相似度较高尚属合理范畴。但这并不代表可以大量复制他人的文章段落组合综述论文,作者除大量逻辑性引用参考文献支持其观点之外,主要应该用自己的原创语言进行阐述和论证,以期达到学科领域内综述之引领作用。

尽管因为体裁不同,论文的相似度的总量有所不同,但是与单篇匹配文献的相似度的要求却是一致的。一般来说,与单篇匹配文献的文本重叠少(如 200 个单词以内)为合理引用借鉴;如出现大段文字重叠相似,或单篇匹配文献相似度超过 5%,或论文总相似度超过 50%,我们应该高度警惕,具体分析被检测文档和高相似度的匹配文献全文,分析判断其引用的合理性。

3.3 时滞效应与多次检测的必要性

另外,工作实践中证明至少两次 CrossCheck 检测是必要的。如图 2 所示,该论文在评审过程中初次检测报告显示为较低相似度,表现为合理正常的引用。然而一个月后在论文正式排版发表之前再次检测,却出现极高的相似度。报告中显示了新的匹配文献,大量的内容重复,单篇文献相似度高达 36%,其原因应该是 CrossCheck 数据库的实时更新。

4 结语

同在一个地球村,全球的学术期刊都不可避免会碰到各种各样的学术不端行为。作为编辑,我们本身同样面临职业

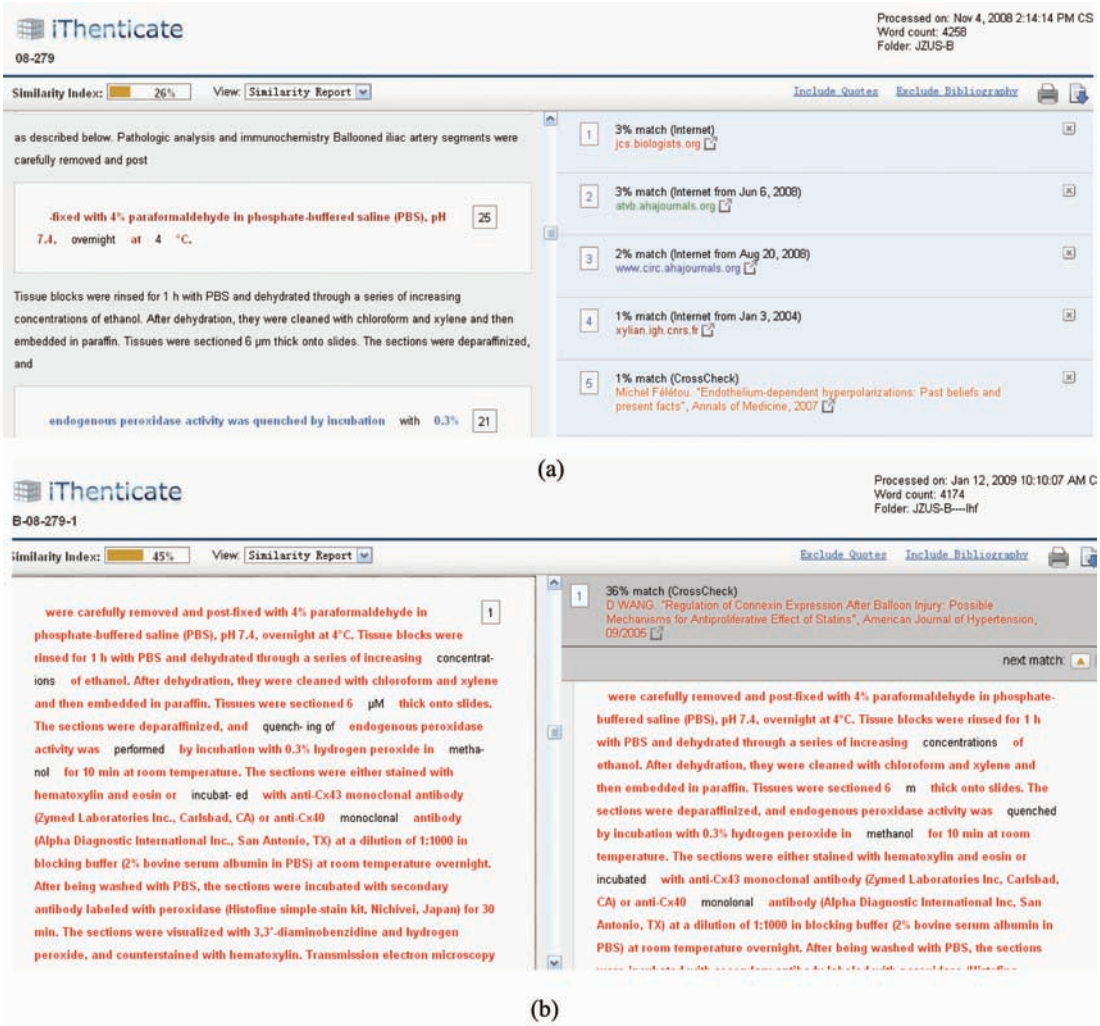

图 2　论文前后两次 CrossCheck 检测的原创性报告结果对比

(a)国际评审中首次检测报告结果;(b)正式发表之前再次检测报告结果

道德的修养和提升,所以在理解学术界不是一块净土的同时,我们有职责去发现和防止这些学术不端行为,并努力将其杜绝在萌芽状态,保护原创作者的著作权,促进学术期刊的健康发展。CrossCheck 作为一个非常有意义的反剽窃工具,尽管存在一些局限和不足,但是其强大的原创性检测功能使其在学术期刊出版工作中发挥着十分重要的作用。

参考文献

1 肖荷露. 美国研究发现中日两国科学家剽窃论文比例略有上升. [2009-03-10]. http://world. people. com. cn/GB/41218/8937987. html

2 Errami M., Garner H. A tale of two citations. *Nature*, 2008, 451: 397 – 399, Published online 23 January 2008 [doi: 10. 1038/451397a]

3 Sorokina D., Gehrke J., Warner S., Ginsparg P. Plagiarism Detection in arXiv. In: Data Mining, 2006. ICDM '06. Sixth International Conference on Publication Date: 18 – 22 Dec. 2006: 1070 – 1075

4 Rampell C. Journals May Soon Use Antiplagiarism Software on Their Authors. The chronicle of higher education. 2008. Arp. 18, http://chronicle. com/free/2008/04/2546n. htm

5 2008 ALPSP Award for publishing innovation-Winner: CrossCheck from the CrossRef/iParadigma Partner. www. alpsp. org

6 http://www. crossref. org/crosscheck. html

7 iThenticcte. [2009-06-09] https://www. ithenticate. com/services. html

8 汪新红,彭绍明. 数字出版平台及其在科技期刊编辑出版工作中的应用. 中国科技期刊研究,2009,20(2): 204 – 213

9 王丹红. 学术欺诈案频发,学术期刊如何应对. 科技时报, 2008-04-02(2)

10 方舟子. 如何避免学术不端行为. [2007-02-15]. http://scitech. people. com. cn/GB/5403070. html

11 哈佛关于"抄袭"的规条.《世界出国》ICXO. COM [2006-01-12]. http://abroad. icxo. com/htmlnews/2006/01/12/754343. htm