# 存在不合理水平文字复制率的稿件作者行为模式及对策分析*

薛婧媛[1)2)] 游 滨[1)] 郭 飞[1)]
1）重庆大学期刊社，400030，重庆；2）重庆大学法学院，400030，重庆

摘 要 文字复制率高的文章不一定涉及抄袭、剽窃，但较为可疑，需要仔细甄别。本文通过分析2本学术期刊特定时间内来稿中文字复制率较高的可疑文档及来源文献的特征，探寻作者的行为模式。结果表明，来源文献主要集中于期刊论文、学位论文；作者倾向于选择新近出版的文献；自然科学稿件涉及的来源文献的文章平均数量比社会科学稿件少。文章根据作者行为模式的特点，提出制定相应的政策和规范，通过规制和引导人的行为来发挥作用，建立健全对学术不端行为的监管和处理机制等对策以防范学术不端，以预防为主，惩罚为辅。

关键词 文字复制率；学术不端；来源文献；学术规范

开放科学（资源服务）标识码（OSID）

在互联网和学术环境大开放的情形下，获取信息和资源的手段非常便利，抄袭、剽窃等学术不端行为变得异常活跃。在全球，这种现象正呈逐步增加的趋势。学术不端行为完全违背了学术规范和道德准则，严重阻碍了科学技术的发展，对学术界乃至整个社会危害极大。学术诚信已成为当前学术界和教育界广泛研究的问题。美国自然科学基金会将捏造、伪造和剽窃认定为学术不端的3种形式[1]。中国教育部于2009年发布《关于严肃处理高等学校学术不端行为的通知》，其中明确指出7种学术不端行为[2]。2016年，中华人民共和国教育部令第40号《高等学校预防与处理学术不端行为办法》发布，更为详细地规定了学术

* 基金项目：国家社会科学基金项目西部项目“学术期刊的数字化转型升级研究”（15XXW001），教育部人文社会科学研究一般项目“高校哲学社会科学期刊数字化建设研究”（14YJAZH100），重庆市社科规划重点项目“传统报刊数字化转型升级研究”（2014ZDCB20）。

D37G86VB00018AOR.html.
[5] 徐明星，田颖，李霁月. 图说区块链：神一样的金融科技与未来社会[M]. 北京：中信出版社，2017.
[6] 卿苏德. 区块链在物联网中的应用[EB/OL].（2017-06-27）[2018-04-21]. http://gngj.gog.cn/system/2017/06/27/015835020.shtml.
[7] 中国电子技术标准化研究院. 中国区块链与物联网融合创新应用蓝皮书[EB/OL].（2017-09-13）[2018-04-01]. http://www.cesi.ac.cn/images/editor/20170913/20170913145041632.pdf.
[8] 他们将区块链技术和人工智能相结合，突破区块链技[EB/OL].（2017-07-28）[2018-04-25]. http://www.sohu.com/a/160476320_684500.
[9] 区块链研习|区块链与人工智能的脑洞时刻[EB/OL].（2017-11-14）[2108-04-23]. http://www.donews.com/technology/detail/15492174.
[10] 当区块链遇上共享经济，未来我们可能一无所有但也能租赁一切[EB/OL].（2016-10-20）[2018-04-25]. http://money.163.com/16/1020/10/C3QJKTHA002580S6.html.
[11] 曾响铃. 绑上区块链的共享经济，玩法大不一样[EB/OL].（2017-07-16）[2018-04-29]. http://www.sohu.com/a/157527964_491065.

（责任编辑：张广萌）

不端的各种行为[3]。

虽然已有一些严格的学术执行规范和约束，但还是有科学家、研究生因为各种原因铤而走险，如完成考核指标、获取研究经费、毕业等。Fang等[4]研究发现PubMed收录的文章因学术不端而撤销的稿件的比例明显高于因科学错误而撤销的稿件，且这一数量在近20年持续增加。JARIĆ [5]指出剽窃和重复发表的文章数量约占出版文章总量的4%。为防止和杜绝学术不端行为的发生，学者、编者、出版商、技术公司积极探索各种行之有效的方法。国际出版伦理委员会（COPE）以流程图清晰明确地提供实施步骤，帮助编辑有效应对学术伦理问题，防止学术剽窃和欺诈[6]。反剽窃软件很大程度地避免了潜在剽窃论文的发表，如英文剽窃检测系统CrossCheck基于检索和文本指纹技术对海量文本快速匹配，以发现可疑片段；中国的CNKI学术不端检测系统是最常用的中文学术论文检测系统，目前提供中英文论文的检测，但以中文为主。

剽窃检测软件广泛运用之前，很少有关于作者行为模式的分析，相关研究大部分集中于采访和调查[7-8]。目前，对剽窃的认定没有明确的界限，维基百科[9]、学术道德[10]、出版商的出版伦理[11]、高校的学术规范[12]、政府机构的学术政策[13]等都有各自的定义。Zhang等[14-15]根据CrossCheck的检测结果来判断剽窃程度，认为文字复制率8.99%±4.23%为轻度剽窃，21.69%±5.65%为中度剽窃，38.78%±10.77%为重度剽窃。IEEE出版服务和产品版操作手册根据文字复制率将剽窃划分为5个层次[16]。一份非正式调查显示，某学术期刊有23%的来稿因涉嫌剽窃而被拒稿[17]。2008年10月至2009年5月，《浙江大学学报》（A&B）大约有22.8%的来稿存在不合理水平复制或者自剽窃现象[14]。《资源科学》通过检测2009—2011年自由来稿，结果显示约1/4的稿件文字复制率≥20%[18]。虽然文字复制率高的文章不一定涉及抄袭、剽窃，但较为可疑，需要仔细甄别。因此，本文分析不同程度的文字复制率、来源文献的特征以探寻作者的行为模式，同时分析作者的行为模式在自科科学类与社会科学类稿件中的主要区别，为编辑、审稿人、出版者更好更快地甄别稿件提供参考，也为分学科制定政策和规范来规制、引导作者的行为提供依据。

## 1 研究对象

本文选取2本中文学术期刊自2011年9月1日至2012年8月31日来稿的原创论文（不包括综述文章）为研究对象，分别代表自然科学类稿件和社会科学类稿件。剽窃检测系统为CNKI的学术不端检测系统的子系统。目前CNKI的剽窃检测对比数据库是中国最全的，包括中国学术期刊网络出版总库、中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库、中国重要会议论文全文数据库、中国重要报纸全文数据库、中国专利全文数据库、互联网资源以及部分英文数据库等。该系统发布于2008年12月，2010年开始被中国各期刊编辑部广泛运用。前述来稿时间段的选取正是基于检测系统广泛运用时间点的考虑。笔者认为该时间段的稿件作者还不十分了解反剽窃软件，不会有意识地替换文字或者重述以逃避检测系统的检测，因此数据反映出来的特征会更明显。同时，笔者在分析时不主观判断复制内容是否涉嫌学术不端，而是客观地以文字复制率作为不合理水平程度的划分依据：<10%为轻度不合理，>20%为中度不合理，>40%为重度不合理，且着重分析中度和重度不合理水平文字复制率的

稿件。

## 2 方法及数据

笔者不主观判断重叠内容在文中的重要程度，仅根据文字复制率进行定性的分析，因为作者自身表述与其他文献大部分完整段落雷同的情况极端偶然且微乎其微[19]。

（1）利用CKNI学术不端检测系统检测2本中文学术期刊自2011年9月1日至2012年8月31日来稿的原创论文，分别为725篇和1 236篇，检测时间为来稿后的1天内。检测结果显示分别有110篇和257篇稿件的文字复制率>20%（排除提前检测的论文。CNKI系统能通过文本指纹技术检测到经过细微修改的再次检测稿件，并标记为“该文献可能被提前检测”以提醒检测者）。可疑文档的文字复制率是指可疑段落所含字数占可疑文档总字数的百分比。系统标记可疑段落为红色，同时提供可疑段落对应的来源文献，每篇来源文献的题目、单篇文字复制率（抄袭于单个来源文献的字数占可疑文档总字数的百分比）、出处、入库时间也会详细标明。

（2）将来源文献的“出处”进行分类：期刊论文、会议论文集、学位论文、网络、报纸、其他（包括专利、标准、法律法规、项目书等）。然后对所有来源文档进行数据清理：① 删除标记为“该文献已经被删除”的条目（原因可能是来源文献被撤稿，或者含有涉密内容被系统屏蔽）；② 删除重复的网络来源文献。成千上万的网站具有无限制的转载功能，造成大量的重复信息，为了避免重复计算复制率，采取只保留一个网络副本的做法。删除标准为单篇复制率相同，且具有相同标题的网络内容，最后得到有效的来源文献数量为962条（自然科学类）和3 466条（社会科学类）。

## 3 结果及讨论

### 3.1 整体情况分析

由于自然科学类和社会科学类的稿件总量不同，为了方便比较，将稿件总量作归一化处理，按照稿件的文字复制率从小到大进行排列（0~100%）。图1为可疑文档文字复制率的趋势情况。可以看出，自然科学类和社会科学类稿件的不合理水平复制的表现大致相同。文字复制率在20%~60%区间内，二者情况几乎一致；在60%~100%区间内，自然科学类显得更为平缓，社科科学类则变得更加陡峭。究其原因，可能是因为自然科学类稿件含有许多图表、公式和计算，而检测系统不能精确地辨别这些内容的复制情况。图1中得到的结果与文献[18]的结果基本一致。图2为可疑文档文字复制率的分布情况，横坐标为可疑文档文字复制率的不同区间，纵坐标为可疑文档数量占总稿件量的百分比（$p_d$）。可以看出，二者的分布规律大致符合幂律分布规律，这与其他自然和社会现象如单词词频分布、引文次数分布、个人经济收入分布、万维网分布等相一致[20]。

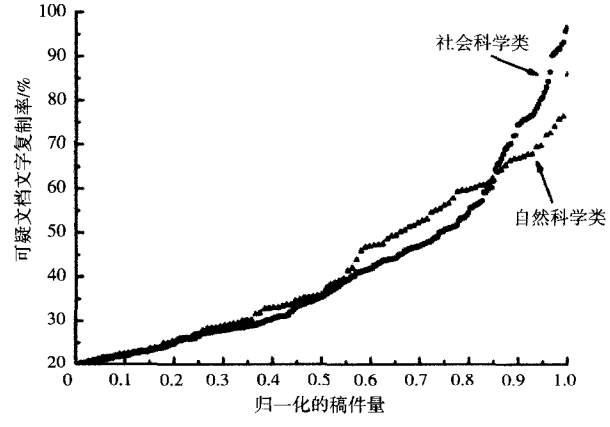


图1 可疑文档文字复制率的趋势情况

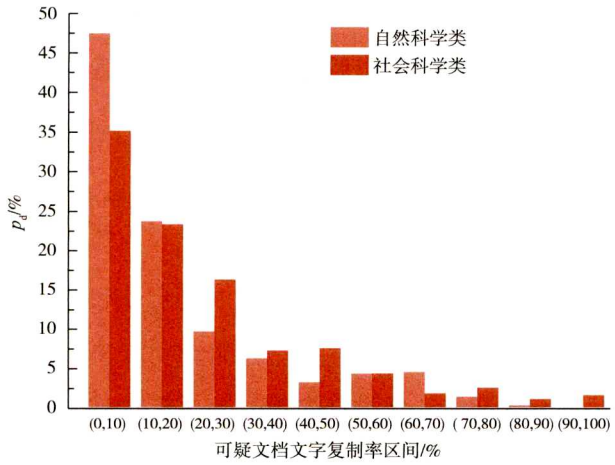


图2 可疑文档文字复制率的分布情况

### 3.2 来源文献的分布情况

表1列出自然科学类和社会科学类稿件的复制来源文献分布。可以看出，两类稿件的主要复制来源都为期刊论文和学位论文。自然科学类稿件中，复制来源为期刊论文的比例（47.81%）与学位论文的比例（48.44%）几乎一样；而社会科学类稿件中，复制来源为期刊论文的比例比学位论文的比例高出35.78%。此外，社会科学类稿件中的网络、报纸来源比例均高于自然科学类稿件。可能在于自然科学类稿件的文字描述中专业术语较多，少见于网络和报纸，而社会科学类稿件的一些文字表达与网络、报纸的用语习惯相近。

从检测结果来看，文字复制率>20%的112篇自然科学类和295篇社会科学类稿件的复制来源文献都不止1篇，自然科学类和社会科学类稿件分别对应的来源文献总数为962条和3 466条，同时系统列出了各来源文献的单篇文字复制率。因此，两类稿件的来源文献篇均数分别为962/110=8.75，3 466/257=13.49。笔者比较了可疑文档的文字复制率与各来源文献的单篇文字复制率总和之间的关系（图3）。可以看出，任何可疑文档的文字复制率总是小于或等于单篇来源文献复制率相加之和，而有的单篇复制率相加之和甚至高达350%以上。这反映出一个不容忽视的问题，来源文献之间存在数量不等的重叠文字，可能涉及抄袭或重复发表。笔者推断这些来源文献可能出版于检测系统发布之前，那时人工很难发现这些重叠文字，而当检测系统发布后，之前的抄袭现象也随之暴露出来。由于来源文献有相互重复的现象，因此真正的来源文献总数应当分别小于962条和3 466条，实际的复制来源文献篇均数也应该小于8.75和13.49。

### 3.3 可疑文档与来源文献的时间关系

可疑文档的作者在选择来源文献时倾向于最

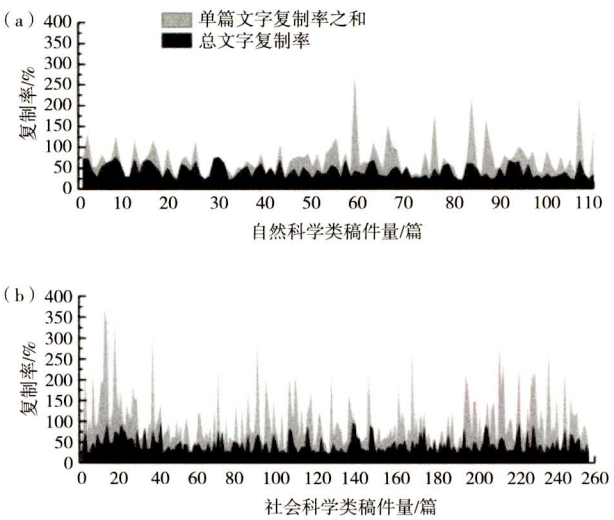


图3 可疑文档的文字复制率与各来源文献的单篇文字复制率总和之间的关系

表1 自然科学类和社会科学类稿件的复制来源分布

| 稿件类别 | 期刊论文 | 会议论文集 | 学位论文 | 网络 | 报纸 | 其他 |
|---|---|---|---|---|---|---|
| 自然科学类 | 460（47.81%） | 21（2.18%） | 466（48.44%） | 5（0.52%） | 2（0.21%） | 8（0.83%） |
| 社会科学类 | 2146（61.92%） | 56（1.62%） | 906（26.14%） | 224（6.46%） | 127（3.66%） | 7（0.20%） |

近出版的还是相对较久的？为此，本文分析了二者的时间关系。由于无法确定作者文章写作的时间，且成稿到投稿之间也可能因作者的个人选择长短不一，因此选择投稿时间作为时间点A，被检测出的来源文献入库时间为时间点B，以A与B的时间差来分析可疑文档与复制来源文献之间的时间关系。因为网络来源文献无法确定准确的发布时间，网络快照时间不能代表首发时间，所以此处仅选取来源文献中的学位论文和期刊论文的入库时间作为比较点，这两类来源文献的数量分别为926条（自然科学类）和3 052条（社会科学类）。图4显示了来源文献数量与时间差之间的关系。可见，来源文献的时间跨度比较大，自然科学类稿件的来源文献的出版时间最近的不到1个月，最久的为1984年；社会科学类稿件的来源文献的出版时间最近的也不到1个月，最久的为1975年。此外，自然科学类稿件的来源文献多为2~3年内发表的文章，发表3年以上的文章呈逐年递减趋势，而社会科学类稿件的来源文献多为1~2年内发表的文章，发表2年以上的文章逐年递减。两类稿件的来源文献为发表10年以上的文章占比很少。

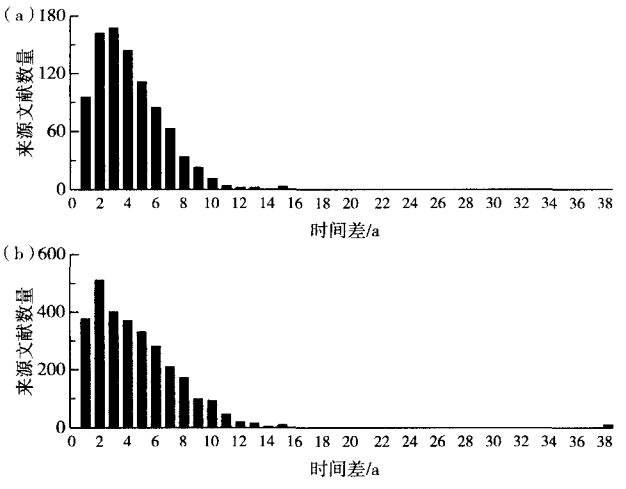


图4 来源文献数量与时间差之间的关系：a)自然科学类，b)社会科学类

总的来说，自然科学类和社科科学类稿件的作者都倾向于选择最近出版的文章作为复制来源，遵循选新不选旧的原则。

### 3.4 来源文献数量偏好

检测系统给出的单篇来源文献复制率是从高到低排列的，如前所述，虽然来源文献可能有相互重叠的现象，但是排名第一的单篇文献肯定为可疑文档复制的最大来源。因此，笔者以最大单篇复制率/总复制率的百分比来判断作者倾向于选择单一来源还是利用多篇文献进行拼凑。图5为最大单篇复制率/总复制率的百分比各区间的可疑文档数量分布情况，即单篇最大占比复制内容的分布情况。横坐标为最大单篇占比的大小（范围0~100%，分为10个区间），数值越小表示来源文献越分散，数值越大表示来源文献越集中，极端90%~100%表示复制内容几乎出自同一篇文献；纵坐标为观测的可疑文档累积量。可以看出，自然科学类稿件在区间90%~100%的数量明显高于其他区间，表明大多数作者倾向于选择1篇文章作为复制来源；社会科学类稿件数量主要集中的区

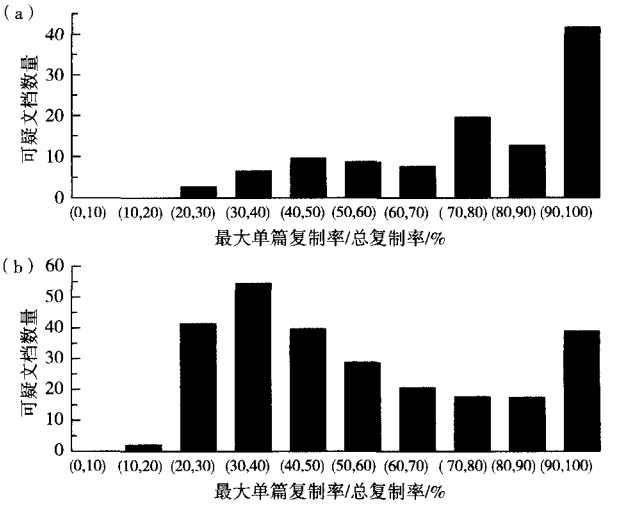


图5 最大单篇复制率/总复制率的百分比各区间的可疑文档数量分布情况：a)自然科学类，b)社会科学类

间为30%~40%和20%~30%，表明大部分作者倾向于利用多篇文献东拼西凑。

## 4 对策及总结

虽然本文属于小样本分析，但仍发现了一些自然科学类和社会科学类可疑稿件作者行为模式的异同：① 两类稿件文字复制率的分布情况基本一致，分布规律大致符合幂律分布，可以初步判断为人类选择行为的结果，属于多个个体的独立行为形成的幂律现象，生成这种幂律分布的具体机制还有待深入研究。② 复制来源文献的总体分布是相同的，集中于期刊论文和学位论文。③ 作者都倾向于选择最近出版的文章作为复制来源，遵循选新不选旧的原则。④ 自然科学类稿件的复制来源文献的平均篇数低于社会科学类。

根据作者行为模式的特点，笔者提出以下对策以防范学术不端。

1）加强学术规范。通过人工比对了部分可疑稿件与来源文献，发现其中有些稿件的文字复制率偏高是由于没有正确地引用他人的学术成果，还有一些是复制了部分自己之前发表的研究成果。因此，编辑部需要向作者强调规范引用，合理利用文献，而不是逐字复制。同时还应加强作者对自我剽窃的认识，了解并避免自我抄袭。

2）提高学术不端检测系统的人工智能程度。借助学术不端检测系统严控不合理的文字复制率是防范学术不端的重要手段，可以从技术层面快速、高效地阻断可疑稿件的发表。但其检测结果只能作为一种参考，有一定局限性。当前的反剽窃检测软件仅限于同一语言体系的字符串匹配检测，虽然能有效地检测逐字逐句的复制以及部分文字的替换和调序，但尚无法检测重述型或者翻译型的文字剽窃，也不能识别图片和公式的复制情况。提高反剽窃检测软件的人工智能程度有赖于信息技术的发展，只有模式识别、检索技术、机器翻译、语义识别等人工智能技术达到一定程度，反剽窃检测软件才能提供更优秀、全面的检测结果。

3）协助审稿专家甄别可疑稿件。文字复制率的高低仅代表文字的重叠，与创新性没有必然联系。因此，可疑文档是否真的涉及学术不端，一定要由相关领域专家将其与来源文献进行仔细比对，对内容进行公正的判断和评价，从而得出客观的结论。编辑部需要将来源文献和可疑稿件一并提交给专家审理，相关领域专家在比对时可以着重于稿件的不同部分以甄别稿件，重点注意是否存在表格、图片和数据的抄袭。

4）多时段检测跟踪论文。学术不端检测系统的论文比对库必然有滞后，可能在投审稿期间不能及时发现文字重复问题，因此建议编辑在文章发表之前再次检测，把好出版环节的最后一关。甚至在发表之后一段时间内继续跟踪，以便尽早发现问题，尽快启动撤销学术不端稿件、消除不利影响等应对措施。

5）建立健全对学术不端行为的监管和处理机制。来源文献作者更容易发现特定论文中不合理复制、替换等学术不端行为，监管部门可以建立合理的举报通道，以便及时发现问题。监管部门还可以制定相应的政策和规范通过规制和引导作者的行为来发挥作用，建立健全对学术不端行为的监管和处理机制，以预防为主，惩罚为辅。

由于数据源有限，本文仅讨论了2本学术期刊投稿作者在特定时间段的行为情况，并不能代表其他自然科学类和社会科学类期刊的作者行为。有研究表明，在剽窃态度上有着明显的文化差异[21]，因此更大范围的作者行为模式分析需要

更深入的研究。任何检测软件都是被动的，并不能消除或完全阻止学术不端行为的发生，只有建立完备的科研诚信体系才能最大限度地减少学术不端行为。

参考文献

[1] National Science Foundation (NSF). NSF's research misconduct regulation [EB/OL]. [2017-5-13]. https://www.nsf.gov/oig/_pdf/cfr/45-CFR-689.pdf.

[2] 中华人民共和国教育部. 教育部关于严肃处理高等学校学术不端行为的通知[EB/OL]. [2017-05-13]. http://www.gov.cn/gzdt/2009-03/21/content_1264527.htm.

[3] 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[EB/OL]. [2017-05-13]. http://www.moe.edu.cn/srcsite/A02/s5911/moe_621/201607/t20160718_272156.html.

[4] FANG F C, STEEN R G, CASADEVALL A. Misconduct accounts for the majority of retracted scientific publications [J]. Proceedings of the National Academy of Sciences of the U.S.A., 2012, 109(42): 17028-17033.

[5] JARIĆ I. High time for a common plagiarism detection system [J]. Scientometrics, 2016, 106(1): 457-459.

[6] COPE. What to do if you suspect redundant (duplicate) publication [EB/OL]. [2017-06-05]. https: //publicationethics.org/files/Full set of English flowcharts_9Nov2016.pdf.

[7] BRUTON S, CHILDERS D. The ethics and politics of policing plagiarism: a qualitative study of faculty views on student plagiarism and Turnitin [J]. Assessment & Evaluation in Higher Education, 2016, 41(2): 316-330.

[8] MAURER H, KAPPE F, ZAKA B. Plagiarism: a survey [J]. Journal of Universal Computer Science, 2006, 12(8): 1050-1084.

[9] Wikipedia. Plagirism[EB/OL]. [2017-06-12]. https://en.wikipedia.org/wiki/Plagiarism.

[10] MARTINSON B C, ANDERSON M S, DE VRIES R. Scientists behaving badly [J]. Nature, 2005, 435(7043): 737-738.

[11] Elsevier. Publishing ethics resource kit (PERK) for editors [EB/OL]. [2017-06-12]. https: //www.elsevier.com/editors/perk.

[12] Harvard University. Harvard plagiarism policy [EB/OL]. [2017-06-12]. http: //isites.harvard.edu/icb/icb.do?keyword=k70847&pageid=icb.page355322.

[13] Office of Science and Technology Policy of the United State. Federal policy on research misconduct [EB/OL]. [2017-06-12]. http: //www.aps.org/policy/statements/upload/federalpolicy.pdf.

[14] ZHANG Y H. CrossCheck: An effective tool for detecting plagiarism [J] Learned Publishing, 2010, 23(1): 9-14.

[15] ZHANG Y H, JIA X Y. A survey on the use of CrossCheck for detecting plagiarism in journal articles [J]. Learned Publishing, 2012, 25(4): 292-307.

[16] IEEE. IEEE publication services and products board operations manual 2016. [EB/OL]. [2017-06-13] http: //www.ieee.org/documents/opsmanual.pdf.

[17] BUTLER D. Journals step up plagiarism policing [J]. Nature, 2010, 466(7303): 167.

[18] 李家永, 耿艳辉, 张戈丽.《资源科学》自由来稿的文字复制状况分析[J]. 中国科技期刊研究, 2012, 23(2): 256-260.

[19] SOROKINA D, GEHRKE J, WARNER S, et al. Plagiarism detection in arXiv [J/OL]. [2017-08-22]. https: //arxiv.org/ftp/cs/papers/0702/0702012.pdf.

[20] BARABÁSI A L, ALBERT R, JEONG H, et al. Power-law distribution of the World Wide Web [J]. Science, 2000, 287(5461): 2115.

[21] EHRICH J, HOWARD S J, MU C, et al. A comparison of Chinese and Australian university students' attitudes towards plagiarism [J]. Studies in Higher Education, 2016, 41(2): 231-246.

(责任编辑：张广萌)