

·科技纵横捭阖·
文/化柏林

抄袭检测系统将给中国学术界带来的变化

学术抄袭是个敏感的话题,加强管理、强调自律等措施发挥着重要作用。然而,利用技术手段解决学术抄袭也是一个重要举措。清华同方(中国知网)研发的学术不端检测系统、武汉大学沈阳副教授研发的 ROST 反剽窃系统、笔者研发的文章自写度测评系统 ANES-SM 等,运用不同的技术手段来解决学术抄袭检测的问题。从自动化角度解决学术不端的问题主要包括两个方面:抄袭检测和一稿多发检测。从技术实现的角度来讲,是一回事,就是相似文献内容的匹配判定。

1 抄袭检测系统将给中国学术界带来的变化

1) 编辑部的审稿工作会得到某种程度的减轻

编辑部利用抄袭检测系统对作者来稿首先进行自写度初查,看看文章有多少成分是自己写的,有多少成分是他人的。如果引用或抄袭他人的成分太多,这样的稿件将不会进入下一轮审稿。而且如果作者有较严重的抄袭的话,那么编辑部可以把此人加入黑名单,以后此作者再投的稿子,就需要慎重考虑甚至不考虑了。

2) 期刊稿源数量将会大幅下降

最近很多编辑部采取缩短出版周期(双月刊变月刊、月刊变半月刊)、增加页码等方式应对来源稿件的增加。笔者对数个研究专题的论文进行集中阅读,在对每个专题看到 20 篇(左右)论文的时候就看不下去了,因为论文之间的重叠度太高了。基于这种检测系统,有问题的论文会被检测出来,而对于作者来讲,由于这种系统的出现,有问题的论文在投稿时也会很小心,所以稿源数量会在一定程度上有所下降。

3) 一稿多发的现象会得到缓解

目前,一稿多发的现象也屡见不鲜。把标题改一下,内容稍作调整,分别投向会议与期刊,或者不同学科的期刊。一稿多投的稿件其论文标题一般会有两个关键词,通过把两个关键词位置关系置换达到“改头换面”的目的,例如《自然语言理解的全信息方法论》与《全信息自然语言理解方法论》。一稿多发由于时间同步,所以尚不好从源头上解决,属于后控操作,



化柏林 中国科学技术信息研究所助理研究员。

栏目主持人 关增建,上海交通大学人文学院常务副院长,教授,兼任中国科学技术史学会副理事长、上海市科技史学会副理事长、《中国科技史杂志》编委、《自然科学史研究》编委。

所以一稿多发的作者进入黑名单也能起到警示与惩罚的作用。

4) 科研诚信评价将会启用自动化手段
一稿多发、学术抄袭检测的自动化,使科研诚信评价制度化。项目评审、职称晋升、博导院士遴选将会采用一票否决制,只要出现较严重的一稿多发或学术抄袭等学术不端行为,那么将会失去以上评选与晋升的机会。科技部、教育部、国家自然科学基金委、新闻出版总署等部门将会联合启动学术抄袭检测大行动。对于这种原则性的问题,学术界本身会更加强调学术规范的自律行为,相关管理机构会加强这方面的监管与处罚。原则性的错误不需要同情也不能同情。

5) 出书周期将会变长

现在有的人几个月就能出本书,更有甚者几周就能出本书。这样的行为称之为“编书”,而不是“写书”或“著书”。编书的自写程度不超过 1/3,除少数章节外,大部分内容来自他人成果。编书的参考文献数量在 10 篇左右,一般不超过百篇,而且以书为多,论文特别是外文参考文献较少。书中不作参考文献的实际标注,却有成页成页的内容来自他人成果。抄袭检测系统的应用,使出版社今后对此类书的出

版会很小心。

2 抄袭检测系统的应用将带来的一些副产品

1) 文献综述将由系统完成

按学术抄袭检测系统的原理,检测出相同的句段,而不同的句段按照一定顺序组合在一起,即可形成综述报告。从理论上讲,如果运用句子匹配分析技术,把相关主题的文章综合到一起,进行句子级的滤重与重组,就可以实现综述型文章的自动完成,即文献自动综述。

2) 参考文献的标注将实现自动化

已有一些参考文献的管理软件,如 EndNote、NoteExpress 等,可以很好地管理并标注参考文献。但是标注在文章的哪个地方,这些软件实现不了,还需要人工标注。

3) 科学计量将会引入引用量等指标

现在的引文分析主要反映引用次数。引用某篇文章的一个术语、一个数据和引用一篇文章的几句话甚至文章的中心论点是不一样的,因此不仅要考虑参考文献与引用文献数量,还要考虑引用内容的量,即不仅要关注引用的多和少,还要考虑引用的大与小,引用粒度的不同决定着参考文献的价值不同。应用抄袭检测系统的原理,可能会使科学性引入引用量的指标。

对于抄袭的认定,不同的研究人员有不同的算法,但基本思路是一致的。抄袭的技术判断主要是段落匹配或句子匹配甚至词语匹配。抄袭检测系统的主要障碍在于没有一家数据库商能提供论文、图书、报纸、专利、网页等十大文献源数据,跨语言匹配也是抄袭检测系统的一个技术难点。

抄袭检测系统会带来上述变化,如果有一天该系统有全文数据库支持和跨语言匹配,抄袭检测系统所带来的变化或许真的很大。

但必须指出这种检测只是形式上的检测,对软性抄袭还无能为力。所谓软性抄袭,是指抄袭人家的观点和论证方式,但用自己的语言写出来,当成自己的创新。这种形式的抄袭,现在的反抄袭软件还判断不出来。

(责任编辑 王芷)