基于逻辑原则的科技论文自动校对方法

■侯修洲 黄延红

收稿日期: 2018-05-09
修回日期: 2018-06-29

《中国科学》杂志社,北京市东城区东黄城根北街 16 号 100717

摘 要 【目的】探索科技论文自动校对方法,以提高编校质量,把编辑从事务性和重复性的劳动中解放出来,使其能够将更多精力投入到创新性的工作中去,适应当前科技期刊出版质量管理日益严格的趋势。【方法】在已有工作的基础上,基于科技论文内在的连续性、一致性和唯一性等全文逻辑原则,在 Word 文档环境中,利用 VBA 辅助编程技术,对论文中出现的不连续文献序号、章节序号、图表序号、公式序号,以及著者-出版年制前后不一致的参考文献进行了识别和标识。此外,还对作者姓名及相应地址和邮编的中英文进行了校对。【结果】VBA 辅助编程技术可识别论文中逻辑相关的编校错误并高亮显示。【结论】在实际工作中,该方法可有效减轻人工校对工作量,减少编校错误,提高出版效率。

关键词 VBA; 逻辑校对; 自动校对; XML
DOI: 10.11946/cjstp.201805090410

科技论文在完成同行评议后,一般还需要经过编辑加工、校对、质检、核对清样等步骤,才能正式发表,这些工作往往比较繁琐,还容易出现差错。2018 年 1 月 10 日,国家新闻出版广电总局报刊司发出《关于对〈报刊质量管理规定〉(征求意见稿)征求意见的通知》,该通知对期刊质量提出了更严格的要求,其中最明显的调整是将期刊编校差错率从 3/10000降低到 2/10000,差错率超过 2/10000 的期刊,其编校质量将被视为不合格。由此可见,编辑的工作压力会越来越大,并将长期陷入事务性的编校工作中,难以发挥编辑的主动性和创造性。

薛子俭等[1]提出分步编校方法,该方法从论文构架核查、分类加工、常规润色、整体核对 4 个方面分步进行,条理清晰,避免了漏校,但所有流程均需要人工参与,并没有减少编辑的工作量,也不能完全保证将每条错误检查出来。近年来,王红剑等[2]和黄城烟[3]提出了利用 Visual Basic for Application (VBA)编程技术在 Word 文档环境中批量替换易错字词的功能,可以在一定程度上降低人工劳动量。龚小谨等[4-6]从自然语言理解和语法分析角度对文章校对进行了研究,该技术的优点是校对颗粒度能达到词语级别,但纠错建议的有效率或首选正确率比较低,与用户的要求还有较大差距。此外,市面上流行的黑马校对软件也是主要集中在词语的错误用法和敏感词的识别,其查错率也有待提高。

近年来,国外大多数期刊均采用了 XML 排版,其优点是论文结构清晰,不仅能为读者提供丰富的阅读体验,而且可从其结构化角度来寻找科技论文内在的逻辑规律,利用这些规律,可对论文进行计算机程序辅助校对。由于 VBA 技术和 Word 文档具有良好的结合性,且笔者已经成功地将 VBA 技术应用于 Word 文档的 XML 结构化标记和参考文献的自动加工中[7-9],在此基础上,本文尝试寻找科技论文的内在连续性、一致性和唯一性等逻辑原则,并基于此原则使用 VBA 辅助编程对科技论文进行自动校对,尽量将大多数错误在排版前标示出来,以提高编排效率,避免多次编校返工。

本研究的编校差错是指排版前可由计算机程序识别的错误。VBA 语言环境、部署及实例应用等内容将不再做详细阐述,可以参考王玥等[10]的文章,语法规则可以参考 http://www.doc88.com/p-931469800915.html。

1 科技论文的连续性、一致性和唯一性原则

科技论文有一定的写作要求,其连续性原则表现在:(1)顺序编码制参考文献著录一般要求正文

基金项目: 2015 年文化产业专项资金项目“中国科技类学术期刊国际传播平台”; 国家自然科学基金“提升我国科技期刊国际影响力的发展战略研究”(71640010)。

作者简介: 侯修洲(ORCID: 0000-0001-5559-3112),硕士,副编审,E-mail: hxz@scichina.org; 黄延红,博士,副编审,总经理助理。

中的文献引用序号必须按照顺序出现，不能漏引；(2) 图表序号、公式序号、章节序号也需要按照顺序出现，不能中断。

一致性原则主要用在著者-出版年制文献的校对。著者-出版年制文献著录一般要求正文中出现的著者年要和文后的文献严格一致。如果正文中著者后面出现“et al.”或“等”的描述，则要求文后文献的作者至少是3位；如果正文中著者后面出现“and”或“和”的描述，则一般要求文后文献的作者是2位。但笔者在实践过程中发现，正文中著者姓的大小写和拉丁文书写格式经常和文后参考文献的著录不一致，如果人工对这部分内容进行校对，其工作量较大，且操作繁琐，并且很难避免疏漏或错误的出现。对于中文科技论文，一致性原则还可以用于校对作者的中文姓名和拼音是否一致，以及中英文地址、邮编是否一致。

无论是顺序编码制文献还是著者-出版年制文献的著录，都要求文后的每一条参考文献只能出现一次，这就是文献的唯一性原则。作者在撰写和修改论文时，由于反复增删内容或其他原因，经常会发生文献重复出现的情况，这时候就需要对文献的唯一性进行检查和校对。

依照上述原则进行校对后，在原文中相应地方进行高亮标识，以提醒加工者注意，这属于建议性质的辅助校对，而不是强制要求用户修改。编辑部可以按照具体体例进行针对性修改，如有特殊情况，可具体问题具体处理。

2 基于连续性原则的自动校对方法

以顺序编码制文献为例，在正文中标注引用文献的格式一般为“[1]”“[1-2]”“[1-3]”“[1-3,5]”“[1-3,5,7,9-11]”等形式，其中的对开线有时也可能为全身线或“~”。首先需要识别这些文献序号，在VBA语言环境中，上述文献格式可以用正则表达式来表述：\[[\d]{1,3}) ((,|.) [\d]{1,3})? \]，其中\[表示开始的方括号，\]表示结束的方括号，([\d]{1,3}) 表示文献序号，((,|.) [\d]{1,3})? 表示结束的文献序号，表示结束的文献序号的“?”也可以省略。如果是像[1-3,5,7,9-11]这样复杂的文献表述，则只需将((,|.) [\d]{1,3})? 在正则表达式中重复几次即可。

当识别了正文中的所有文献序号后，就要判断序号的连续性了。将某一处的文献序号表述内容记为I，将I处之前的文献序号表述内容记为$I-1$，设定$I-1$处的最大文献序号为M_{max}，显然，当程序开始执行时，M_{max}的初始赋值为1。当程序执行到第I处时，求取该处文献序号的最大值和最小值，分别记为I_{max}和I_{min}，此时判断第I处文献序号是否和第$I-1$处文献连续，可以分为三种情况：(1) 当$I_{max} \leqslant M_{max}$时，则$I$处文献和$I-1$处文献连续；(2) 当$I_{min} > M_{max}$时，则$I$处文献和$I-1$处文献不连续，此时将$M_{max}$重新赋值为$I_{max}$；(3) 当$I_{min} \leqslant M_{max} < I_{max}$时，此时则判断$M_{max} \sim I_{max}$之间的每一个数是否在第$I$处文献序号内容中包含，如果包含，判断为连续，否则，判断为不连续，同时将M_{max}重新赋值为I_{max}。顺序编码制文献连续性校对流程图见图1。

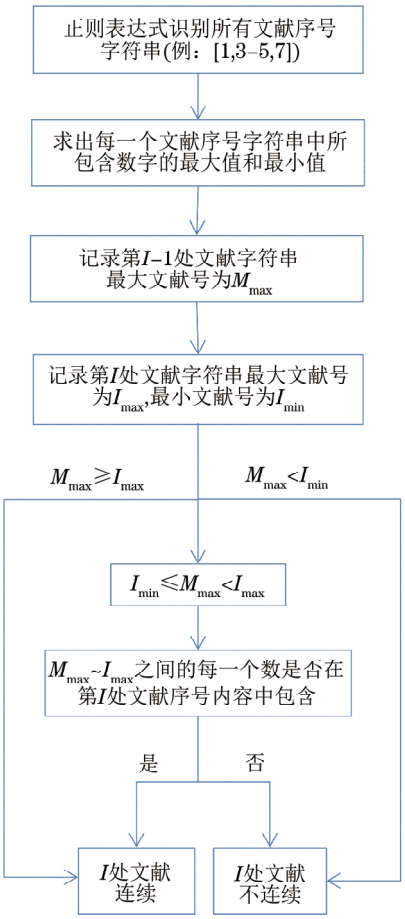
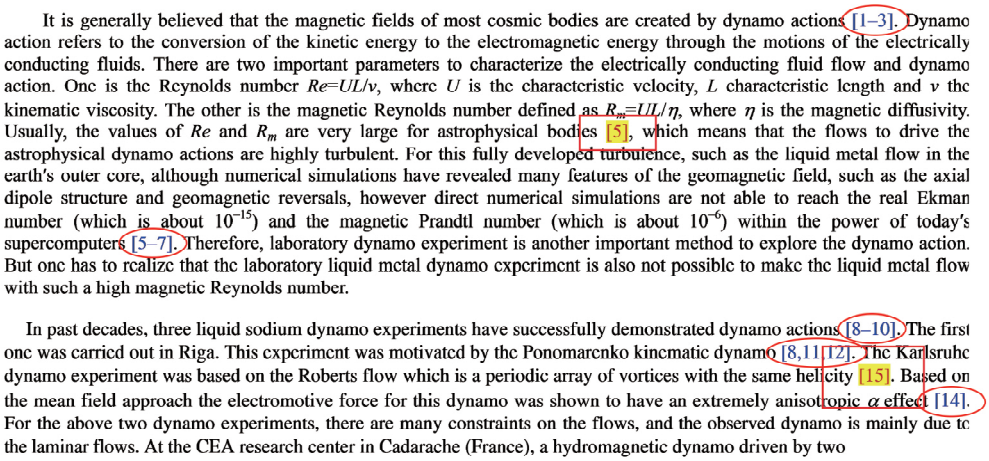


图1 顺序编码制文献连续性校对流程图

在程序运行中，将每一处连续的文献序号标为蓝色，如校对示意图(图2)的圆圈所示；不连续的文献序号标为红色字体并高亮，如图2的方框所示。

图表序号和公式序号的判断规则与顺序编码制文献序号连续性判断规则相同，此处不再赘述。

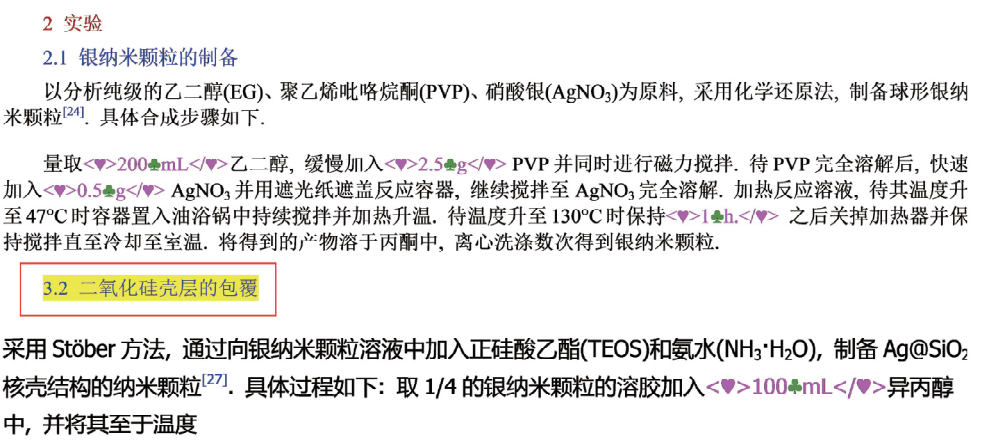
对于章节标题序号的连续性判断，则需要事先定位章节标题的位置，具体标记方法可以查阅文献[7]。


It is generally believed that the magnetic fields of most cosmic bodies are created by dynamo actions [1–3]. Dynamo action refers to the conversion of the kinetic energy to the electromagnetic energy through the motions of the electrically conducting fluids. There are two important parameters to characterize the electrically conducting fluid flow and dynamo action. One is the Reynolds number $Re=UL/\nu$, where U is the characteristic velocity, L characteristic length and ν the kinematic viscosity. The other is the magnetic Reynolds number defined as $R_m=UL/\eta$, where η is the magnetic diffusivity. Usually, the values of Re and R_m are very large for astrophysical bodies [5], which means that the flows to drive the astrophysical dynamo actions are highly turbulent. For this fully developed turbulence, such as the liquid metal flow in the earth's outer core, although numerical simulations have revealed many features of the geomagnetic field, such as the axial dipole structure and geomagnetic reversals, however direct numerical simulations are not able to reach the real Ekman number (which is about 10^{-15}) and the magnetic Prandtl number (which is about 10^{-6}) within the power of today's supercomputers [5–7]. Therefore, laboratory dynamo experiment is another important method to explore the dynamo action. But one has to realize that the laboratory liquid metal dynamo experiment is also not possible to make the liquid metal flow with such a high magnetic Reynolds number.

In past decades, three liquid sodium dynamo experiments have successfully demonstrated dynamo actions [8–10]. The first one was carried out in Riga. This experiment was motivated by the Ponomarenko kinematic dynamo [8,11,12]. The Karlsruhe dynamo experiment was based on the Roberts flow which is a periodic array of vortices with the same helicity [15]. Based on the mean field approach the electromotive force for this dynamo was shown to have an extremely anisotropic α effect [14]. For the above two dynamo experiments, there are many constraints on the flows, and the observed dynamo is mainly due to the laminar flows. At the CEA research center in Cadarache (France), a hydromagnetic dynamo driven by two

图 2　顺序编码制文献序号连续性校对示意图

对于一级标题，只需提取标题前面的序号，按照自然数来判断是否连续，而二级标题和三级标题的序号连续性判断，则不能简单套用自然数来判断。一般二级标题序号为“1.1、1.2、1.3”“2.1、2.2、2.3”等形式，三级标题序号为“1.1.1、1.1.2、1.1.3”“2.1.1、2.1.2、2.1.3”等形式。关于二级标题和三级标题，当成功提取标题序号后，应先忽略序号中的点，然后比较自然数顺序序列。与判断一级标题序号连续性不同的是，当考虑二级标题序号的连续性时，既要满足自然数连续性规则，也要保持该二级标题序号的第一位数和紧邻的一级标题序号一致；当考虑三级标题序号的连续性时，同样要考虑该三级标题序号的前两位数与紧邻的二级标题序号一致。对于不连续的章节标题，可用高亮显示，如图 3 的方框所示。


2 实验

2.1 银纳米颗粒的制备

以分析纯级的乙二醇(EG)、聚乙烯吡咯烷酮(PVP)、硝酸银($AgNO_3$)为原料，采用化学还原法，制备球形银纳米颗粒[24]. 具体合成步骤如下.

量取<♥>200♣mL</♥>乙二醇，缓慢加入<♥>2.5♣g</♥> PVP 并同时进行磁力搅拌. 待 PVP 完全溶解后，快速加入<♥>0.5♣g</♥> $AgNO_3$ 并用遮光纸遮盖反应容器，继续搅拌至 $AgNO_3$ 完全溶解. 加热反应溶液，待其温度升至 47°C 时容器置入油浴锅中持续搅拌并加热升温. 待温度升至 130°C 时保持<♥>1♣h.</♥> 之后关掉加热器并保持搅拌直至冷却至室温. 将得到的产物溶于丙酮中，离心洗涤数次得到银纳米颗粒.

3.2 二氧化硅壳层的包覆

采用 Stöber 方法，通过向银纳米颗粒溶液中加入正硅酸乙酯(TEOS)和氨水($NH_3\cdot H_2O$)，制备 $Ag@SiO_2$ 核壳结构的纳米颗粒[27]. 具体过程如下：取 1/4 的银纳米颗粒的溶胶加入<♥>100♣mL</♥>异丙醇中，并将其至于温度

图 3　章节序号不连续的示意图

3　基于一致性原则的自动校对方法

一致性校对主要涉及到著者-出版年制文献的校对，一般此类文献在正文中引用时，其表述方式为“姓，年”“姓 et al/等，年”“姓 1 and/和 姓 2，年”“姓(年)”“姓 et al/等(年)”和“姓 1 and/和 姓 2(年)”等形式。基于以上格式，笔者编写了识别著者年的正则表达式：

((\b[a-zA-Z\u00C1-\u00FF\u2C60-\u2C74\u002D]+\b((and|和)\b[a-zA-Z\u00C1-\u00FF\u2C60-\u2C74\u002D]+\b)?)|([\u4e00-\u9fa5]{2,3}(和([\u4e00-\u9fa5]{2,3}))?))(等人|等|et al.|et al|)?(,)?()?(\()?((20|19|18)([\d]{2}))([a-g])?(\))?

当完成正文的著者年信息识别后，还需要将每一条的识别内容和文后参考文献进行比较，其流程如图 4 所示。基于文献[8-9]，笔者已经成功地将参考文献进行了自动加工和 XML 标记拆分，绝大多数参考文献都实现了姓名、文题、刊名、年、卷、页码等信息的拆分(图 5)。只需将正文中识别的姓和年与文后已经拆分的文献信息中的姓和年进行匹配比较即可。如果前后验证没问题，则标上蓝色；如果前

后不对应，则高亮并用红色字体标识。与图 5 文献对应的校对示例如图 6 所示。

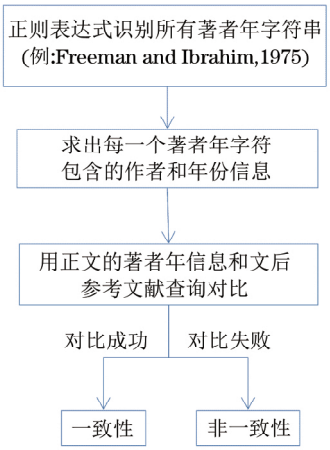


图 4　著者-出版年制文献一致性校对流程

1. <r>\author{<a><g>J</g><s>Colwell</s><a><g>A</g><s>Gulbis</s><a><g>M</g><s>Horanyi</s><a><g>S</g><s>Robertson</s>}, \title{Dust transport in photoelectron layers and the formation of dust ponds on Eros}, \journal{Icarus}, \year{2005}, \vol{175}, \fpage{159}, \lpage{169}, \doi{10.1016/j.icarus.2004.11.001}, \ads{2005Icar..175..159C}</r>

2. <r>\author{<a><g>J W</g><s>Freeman</s><a><g>M</g><s>Ibrahim</s>}, \title{Lunar electric fields, surface potential and associated plasma sheaths}, \journal{The Moon}, \year{1975}, \vol{14}, \fpage{103}, \lpage{114}, \doi{10.1007/BF00562976}, \ads{1975Moon...14..103F}</r>

图 5　参考文献 XML 标记加工示意图[6]

dust particle number density, respectively, then modifications caused by charged dust particles to the sheath potential and plasma profiles would be negligible. In this case, the sheath is referred to as "tenuous" (Nitter et al., 1994). A good deal of research has been conducted on this tenuous sheath, determining the electric field and plasma using a sheath model that enables the charging and dynamic processes of dust to be traced. In this way, basic properties of dust particle dynamics, including oscillation, damping, stability, and trapping, can be investigated (e.g., Nitter et al., 1992; Colwells et al., 2005).

The terminator region on the Moon is the place where surface potential switches from positive to negative (Freeman and Ibrahims, 1975). Horizontal electric field is thought to exist in the terminator region, which would function to launch dust particles over long distances. However, due to a lack of observational data, uncertainties remain about the nature of the plasma environment that extends from the terminator region deep into the lunar wake. To date, observations and research on the lunar sheath and dust dynamics have been limited in most cases to the dayside, and in particular have focused on the subsolar region (Nitter et al., 1998; Poppe and Horányi, 2010). As a result, quantitative estimates of the global properties of the lunar sheath and dust dynamics are currently lacking.

注：第1条文献作者Colwell多了s，第2条文献作者Ibrahim多了s。

图 6　著者-出版年制参考文献一致性校对示意图

对于中文版论文，笔者已经利用 VBA 程序将文档中作者姓名及其相应地址的中英文内容进行了标记[7]，可方便提取每一位作者及其相应地址的中英文信息。将作者的中文姓名转换成拼音，然后和作者标注的拼音进行匹配比较，如果表述不一致则标黄色高亮；同理，可提取地址和邮编的中英文信息，如果不一致，同样黄色高亮提示(图 7)。如果作者及其相应地址的中英文数量不一致，将弹窗提醒加工人员注意。

基于 HSV 颜色空间的水下物体识别技术研究
李代林*
中国石油大学(华东)理学院，青岛 266581;
A new technique for recognition of underwater object based on HSV color space
LI DaiLing*
China University of Petroleum (east China), College of Science, Qingdao 266580, China

图 7　作者姓名和邮编中英文校对示意图

4　基于参考文献唯一性原则的校对方法

笔者已经将参考文献进行了 XML 拆分[8-9]，并且获取了文献的 DOI 信息(图 5)。在提取每条参考文献的 DOI 信息时，如果发现参考文献的 DOI 信息相同，则判断这些文献是重复文献，程序会将重复文献都标为红色字体，具体示例如图 8 的方框所示。

1) <r>\author{<a><g>C.</g><s>Gissinger</s><a><g>A.</g><s>Iskakov</s><a><g>S.</g><s>Fauve</s><a><g>E.</g><s>Dormy</s>}, \journal{Europhys. Lett.}, \year{2008}, \vol{82}, \fpage{29001}, \doi{10.1209/0295-5075/82/29001}, \ads{2008EL.....82.29001G}</r>

2) <r>\author{<a><g>X.</g><s>Wei</s><a><g>A.</g><s>Jackson</s><a><g>R.</g><s>Holzerbach</s>}, \journal{Geophys. Astrophys. Fluid Dyn.}, \year{2012}, \vol{106}, \fpage{681}, \doi{10.1080/03091929.2011.620569}</r>

3) <r>\author{<a><g>C.</g><s>Gissinger</s><a><g>A.</g><s>Iskakov</s><a><g>S.</g><s>Fauve</s><a><g>E.</g><s>Dormy</s>}, \journal{Europhys. Lett.}, \year{2008}, \vol{82}, \fpage{29001}, \doi{10.1209/0295-5075/82/29001}, \ads{2008EL.....82.29001G}</r>

注：文献1)和3)的DOI信息相同，视为重复文献。

图 8　参考文献唯一性校对示意图

5　结论

在已有工作的基础上[7-9]，寻找论文逻辑的连续性、一致性和唯一性原则，并利用这些原则对科技论文进行全文自动校对，不仅减轻了加工人员的劳动量，减少了审校的轮次，也避免了低级编校错误的出现，提高了出版效率，同时也为后期 Word 文档转换为 XML 文件提供了质量保证。

论文的逻辑原则包括：(1) 连续性原则，适用于论文中不连续的文献、章节、图表、公式等序号的校对；(2) 一致性原则，适用于著者-出版年制文献前后不一致表述的校对和作者姓名及相应地址、邮编的中英文校对；(3) 唯一性原则，适用于文后重复出现的参考文献的校对。

需要说明的是，这些原则是笔者在《中国科学》系列刊物编校实践中总结出来的，依照上述原则进行校对的结果，是建议性质的辅助校对，相关编辑部可以按照具体情况进行针对性修改。论文自动校对大致包含逻辑校对和语法校对两个方向，本研究侧重于逻辑原则，笔者也很期待未来能开发出全面且性能良好的校对方法。

参考文献

[1] 薛子俭，付利. 科技论文分步编校法及注意事项[J]. 中国科技期刊研究，2012，23(2)：325-328.
[2] 王红剑，高爱英，游苏宁. 利用 Word 进行自动编校[J]. 中国科技期刊研究，2009，20(3)：502-503.
[3] 黄城烟. 基于 Word 宏技术的易错词自动检索与校对[J]. 编辑学报，2014，26(4)：356-358.
[4] 龚小谨，罗振声，骆卫华. 中文文本自动校对中的语法错误检查[J]. 计算机工程与应用，2003，39(8)：98-100.
[5] 朱磊. 自然语言处理之汉语文本自动校对[D]. 成都：电子科

技大学,2005.
[6] 张仰森,俞士汶. 文本自动校对技术研究综述[J]. 计算机应用研究,2006,23(6):8-12.
[7] 侯修洲,黄延红. 基于 VBA 的 Word 文档 XML 结构化标记方法[J]. 编辑学报,2017,29(5):471-474.
[8] 侯修洲,黄延红. 利用 VBA 程序和 HTTPS 协议获取参考文献的 DOI 信息[J]. 编辑学报,2016,28(5):466-469.
[9] 侯修洲,黄延红. 基于 CrossRef 数据库的参考文献自动加工及 XML 标引方法[J]. 编辑学报,2017,29(1):70-72.
[10] 王玥,毛善锋,刘谦. Word 文档中通过 CrossRef 自动查询与整合英文参考文献 DOI 的实践[J]. 中国科技期刊研究,2013,24(2):333-337.

作者贡献声明:
侯修洲:整理数据,撰写论文;
黄延红:参与论文研究思路的设计,修改论文。

An automatic proofreading method for scientific papers based on the logic principles

HOU Xiuzhou, HUANG Yanhong

Science China Press, 16 Donghuangchenggen North Street, Dongcheng District, Beijing 100717, China

Abstract **[Purposes]** This paper aims to explore the methods of automatic proofreading of scientific papers to improve the quality of editing, and liberate editors from the transactional and repetitive labor, so that editors can devote more energy to innovative work and adapt to the increasingly strict trend of the quality management of scientific journals. **[Methods]** Based on the our existing work, we found the inherent logic principles, such as continuity, consistency, and uniqueness in scientific papers. In the Word document environment, we had found and marked the discontinuous serial numbers about reference, figure, table, formula, and highlights that appear in the text using the VBA assisted programming techniques. The inconsistent expressions of the reference as author-year in the text were identified and displayed. In addition, the author's name, address, and postal code were proofread in both English and Chinese. **[Findings]** VBA assisted programming techniques can identify logically relevant editing errors in the paper and highlight them. **[Conclusions]** In practical work, the proposed method can effectively reduce the workload of manual proofreading, reduce the editing errors, and improve the publishing efficiency.

Keywords: VBA program; Logical proofreading; Automatic proofreading; XML

(本文责编:刘晶晶)