翻訳記事

論文の著作権を守る
学術出版における剽窃検知

Credit where credit's due

Plagiarism screening in scholarly publishing

MEDDINGS, Kirsty¹

.....

¹ Product Manager, CrossRef (40 Salem Street Lynnfield, MA 01940, USA)情報管理 53(3), 140-144, doi: 10.1241/johokanri.53.140 (<http://dx.doi.org/10.1241/johokanri.53.140>)

著者抄録

剽窃は学術出版においては目新しい問題ではないが、インターネットとデジタル出版の時代を迎え、ますます大きな問題となりつつある。こうした時代では、発表論文が単に増加したというだけでなく、オンライン・データベースや出版社の Web サイト、機関リポジトリからの文献の情報検索がかつてないほど容易になった。しかし、コンテンツの数とその利用可能性が増大するにつれ、査読者は、専門分野の全資料を読むことが難しくなり、オリジナルでないコンテンツの検出が一段と困難になっている。この問題に対応するため、CrossRef は学術出版社向けの剽窃検知サービス CrossCheck (<http://www.crossref.org/crosscheck.html>) の提供を開始した。

キーワード

剽窃, 論文, スクリーニング, CrossRef, CrossCheck, 出版社, 査読

1. 問題化する剽窃

学術出版における研究論文の剽窃の程度を推定するため、これまで多くの調査が実施されてきた。米国国立衛生研究所 (National Institutes of Health) の資金提供を受けた数千人の若手・中堅科学者を対象とした2002年の調査の結果、回答者の1.4%が「原著者の許可を取得することなく、あるいは適切な典拠

を表示することなく、他人のアイデアを使用した」と認めており、また4.7%が「同一のデータまたは成果を複数の刊行物に発表した」と告白したことが明らかになった。非回答による偏りと過少報告を考慮に入れると、このような「不正行為の推定件数は控えめである」とこの調査報告の作成者は述べている¹⁾。

2006年に実施された調査は、米国厚生省研究公正局に報告された研究不正件数は「研究機関の調査で

本稿は、『Learned Publishing』誌23巻1号(2010年1月号)に掲載された論文を、著者の許可を得て編集事務局にて翻訳・転載したものである。原文は以下のアドレスから有料で入手できる。<http://dx.doi.org/10.1087/20100102>

は年間平均24件という低い状況」にとどまっているとしている。しかし、調査対象の科学者2,000名超のうち192名（8.7%）が、過去3年間に1件以上の研究不正疑惑のある行為を自分の部門の研究者が行っていることを目撃したり、その確証を持ったりしていると述べている。この192名の科学者は合計265件の事例を挙げており²⁾、その36%は剽窃疑惑であった。この調査報告書は「憂うべき過少申告の実態」を示していた。

3番目の調査はサウスウェスタン・テキサス大学(UT Southwestern)でハロルド・ガーナーのチームが実施したMEDLINEの最新コンテンツの分析である。この調査結果は刊行物やブログで紹介・報告されている。MEDLINEの収録論文を相互に比較した結果、「引用文献は酷似するも、著者の重複はない9,120件」の論文が抽出・データベース化された。2009年2月時点で、「全文解析によって、剽窃の可能性のある212組の論文が特定されている」³⁾。これらの論文は、査読を経て発表されたものである。その後の調査の結果、論文が撤回されるに至った例が数件報告された。

こうした調査より「学術出版には剽窃問題があり、それは過少報告・検知不足の問題である」というのが一致した見方であろう。論文数は常に増大しているが、剽窃との闘いで編集者を支援する技術も利用できるようになっている。

2. 信頼に値する電子コンテンツ

CrossRefの綱領には、「持続可能なインフラの共同開発・利用を推進することにより、信頼できる電子コンテンツを容易に特定し利用できるようにすること」とある。「信頼できる」コンテンツであることの一つは、その典拠があるということである。すなわち、いつ、どこで、誰によってその研究が行われ発表されたかがわかり、他の情報源から引用した文章やアイデアに対して適正な出典表記をしていることでもある。2006年の夏にCrossRefの理事会と会員が剽窃

に関する懸念を表明したとき、CrossRefが剽窃の検知活動を進める組織であるべきだとされたのは当然のことと思われた。その理由は、CrossRefには上述の信頼性を推進するように付託された権限があるというのが一つ、もう一つは、CrossRefは個々の出版社ではできない学術コンテンツの改善を総合的に実施している出版社関連団体であり、CrossCheckのような包括的で効果的なコンテンツ検査システムの実施は、多くの機関が参加してのみ可能となるからである。

剽窃例を探するには、学術コンテンツ同士を比較する必要がある。学生ならWikipediaからコピーしたいと思うだろうが、研究者は同分野の研究者の論文を不正流用する可能性が高い。これまで、大量の既発表論文に対し投稿論文を効果的に検査する剽窃検知手法がなかった。なぜなら、十分なコンテンツに索引付けし真に広範囲の比較ができるようにするには、数百もの出版社との相互協定が必要となり、これは簡単ではないからである。

出版社は、自社刊行物の全文が索引付けされCrossCheckデータベースに加えられることに同意のうえ、CrossCheckに参加する。このデータベースはCrossCheckの全会員に提供され、自社のコンテンツや投稿原稿との比較が可能となる。本稿執筆時点で、CrossCheckデータベースは、約30,000種のジャーナル・書籍・会議録の論文と単行本の章2,200万件を登録している。この数字は、今日までにCrossRefに登録された全コンテンツの半数を超えるものであり、毎週増加している。これらの刊行物はすべての専門分野をカバーしており、アーカイブ分も可能な限り索引付けを行っている。このデータベースはCrossCheck独自のものであり、剽窃検知の成功に極めて重要である。

3. 剽窃検知技術

出版社は、本データベースを利用して文書をスクリーニングする際、CrossRefの技術提携企業の

一社 “iParadigms” 製の確立された剽窃検知技術 “iThenticate” を使用する⁴⁾。同社は、学生の論文スクリーニング用として大学で広く使用されている “Turnitin” システムで最もよく知られており、その基盤技術がiThenticateにも用いられている。このWebベースのiThenticateシステムで文書のアップロードと分析、他のコンテンツとの比較ができる。CrossCheckの場合、比較対象となるコンテンツは上述のデータベースと、PubMedやarXiv.orgなどのコンテンツを含む数百万Webページの最新版とアーカイブ、およびCengageやEBSCOのような業者から提供される商用データベースのコンテンツである。

スクリーニング対象の文書は、分割、分析され、膨大なデータベースと比較され同一または類似の文節が検出される。全体のテキスト類似率が、例えば30%あったとして、この類似率の内訳が複数の論文ごとに、ある文献とは5%、別の文献とは7%一致するなど示される。このシステムは、判定結果を微調整し、引用文献の一致や重要でないと思われる微小の一致を除外するなどの調整ができる。

4. 判断は人

CrossCheckの根幹をなす技術は高度でデータベースは包括的であるが、この強力な組み合わせをもってしても、コンピュータだけで剽窃を検知することはできない。コンピュータはテキストが一致したという事実を示すのみであり、一致した理由の判断を下すことはできない。同じ表現が2か所で現れる妥当な理由として、引用文献一覧、既発表論文からの引用、著者の個人Webサイト上の自分の原稿との完全一致などがある。剽窃が起こったかどうかを判断するためには、当該分野の専門家が、一致率が高かった原稿を添付した報告書を検討することが非常に重要である。

この時点では、「一致率が高い要因は何か」が一番の疑問であるが、それは分野間で異なり個々の出版

物間でさえ異なるというのが答えである。本稿執筆時点において、CrossCheckに参加している出版社の多くは、社内でこのサービスを試用中であり、この間に集めたデータを基に方針を立てるものと思われるが、われわれは現在、詳細調査に着手する類似率は10%から40%までの間に設定されと考えている。

5. 編集プロセスでのCrossCheckの利用

CrossCheck剽窃検知システムを使用し始めようとする出版社が最もよく挙げる問題は、チェックの対象とその時期である。そこには多くの選択肢があり、それぞれに賛否両論がある。編集プロセス開始時に、受け付けたすべての原稿をチェックするようにした出版社もあれば、投稿原稿2または3に対し1原稿などの割合でランダムにスクリーニングする「抜き取り検査」方式を採った出版社もある（ここで重要なのは、ランダム方式は本当にランダムでなければならないということである。例えば、特定国からの文書はすべて日常的にスクリーニングし、他の国からのものはほとんどしなれば、サンプリングにバイアスをかけていると思われる恐れが生じる）。他方、編集プロセスがかなり進んだ時点で原稿をチェックし、採択直前に全文書をスクリーニングしているCrossCheck参加出版社も多い。

後者の方式を採ればCrossCheckシステムで処理する文書量は当然少なく、チェック費用は少なくて済む^{注1)}。が、この方式に対する反対論としては、受付時であれば疑わしいと思った原稿は、懸念の箇所を明らかにして法的・懲罰的に介入することなく却下できる。一方、論文が査読を経て出版に向かう頃には、その論文はパブリックドメインにかなり近い状態にあり、見つけた不正行為は正式に取り扱う必要があるため、費用と時間が余計にかかることが挙げられている。両方式はいずれもメリットとデメリットがあり、本プロジェクトが開始から日の浅い現段階では、参加出版社に特定パターンは現れていない。

同様に、WebベースのiThenticateを単体で使用している出版社があれば、編集工程管理システムに組み込んで活用している出版社もある。この方法には、編集スタッフが一つのシステムを離れ他のシステムを開く必要がないという明らかなメリットがあり、既存の作業フローの一部として原稿を処理することができる。が、それは義務的なものではなく、すべての利用者は、希望すればiThenticateインターフェースからCrossCheckシステムに直接アクセスできる。

6. 剽窃への対応

剽窃の定義方法や剽窃疑惑時の対応方法など微妙な分野を出版社に指導している情報源は数多い。出版倫理委員会（Committee on Publication Ethics）には、状況に応じた行動指針を提言する一連のフローチャートがある⁵⁾。米国研究公正局（Office of

Research Integrity）はWebサイトに剽窃の定義⁶⁾と実質的な研究不正対応指針⁷⁾を載せている。科学編集者会議の「科学雑誌の倫理向上に関する白書」⁸⁾も、剽窃を定義し対処方法を示している。また、世界各国・地域の剽窃問題対応機関も示している。

最後に、CrossCheck活動で協力し合うことにより、出版社は剽窃問題への認識を高め、同僚の論文を不正流用する少数の心ない研究者に対し「剽窃は許されない」というメッセージを強化発信できることを期待したい。剽窃検知プロセスの説明や著者向けの明確な指針、Webサイトや刊行物上のCrossCheckロゴの目立つ表示は、研究者教育となり、疑わしいコンテンツの作成・提出を抑制するのに役立つだろう。大多数の誠実な研究者は、価値あるコンテンツゆえ信頼をおく刊行物が厳格な公正さを維持するためにベストを尽くしていると安心できる。

本文の注

注1) CrossCheckの年会費は、CrossRefの出版社年会費に20%を上乗せすることで包含される（CrossRef年会費は次のアドレスを参照。http://www.crossref.org/02publishers/20pub_fees.html）。また1論文あたりのチェックにかかる料金は0.75米ドルで、論文数が増えると割引価格が適用される。

参考文献

- 1) Martinson, B.C.; Anderson, M.S.; de Vries, R. Scientists behaving badly. *Nature*. 2005, no.435, p. 737-738.
- 2) Titus, Sandra L.; Wells, James A.; Rhoades, Lawrence J. Repairing research integrity. *Nature*. 2008, no.453, p. 980-982.
- 3) Long, Tara C.; Errami, Mounir; George, Angela C.; Sun, Zhaohui; Garner, Harold R. Responding to Possible Plagiarism. *Science*. 2009, vol. 323, no. 5919, p. 1293-1294.
- 4) iParadigms, LLC. <http://www.iparadigms.com>, (accessed 2010-04-23).
- 5) "Flowcharts". Committee on Publication Ethics. <http://publicationethics.org/flowcharts>, (accessed 2010-04-23).
- 6) "ORI Policy on Plagiarism". Office of Research Integrity. <http://ori.dhhs.gov/policies/plagiarism.shtml>, (accessed 2010-04-23).
- 7) "Handling Misconduct: Introduction". Office of Research Integrity. <http://ori.dhhs.gov/misconduct/>, (accessed 2010-04-23).

- 8) "Editorial Policies: Promoting Integrity in Scientific Journals". Council of Science Editors. http://www.councilscienceeditors.org/editorial_policies/white_paper.cfm, (accessed 2010-04-23).

Author Abstract

Plagiarism is not a new problem in academic publishing, but in the age of the Internet and digital publishing it would appear that it may be a growing one. Not only is there more literature available, but it has never been easier to search online database, publisher websites, and institutional repositories. And as the content and its availability grows, so the ability of the reviewer to have read everything in their field diminishes, and detecting unoriginal content becomes harder. In response to this problem, CrossRef has launched CrossCheck (<http://www.crossref.org/crosscheck.html>), a plagiarism detection service for scholarly publishers.

Key words

plagiarism, article, screening, CrossRef, CrossCheck, publisher, peer review