In-depth argument on information

情報論議

根掘り

葉掘り

剽窃の自動検出

名和小太郎

情報管理 53(8), 465-466, doi: 10.1241/johokanri.53.465 (<http://dx.doi.org/10.1241/johokanri.53.465>)

剽窃という行為は、インターネットの普及とともに急激に拡大した。テキストや図表のコピー＆ペーストが簡単になったからである。この現象は、とくに学生のあいだで日常化している。

このコピー＆ペーストであるが、その行為は学生にとどまるのか、研究者は無縁なのか。こんな疑義が当然ながら生じた。その一つとして『ネイチャー』は2005年5月19日号のニュース欄に「不正を駆逐する」という報告を載せた。

この記事の記者は、まず剽窃の実態がよく見えない、バイオメディカル分野に0.02%、麻酔分野に5%、外科分野に22%という調査はあるが、と示している。

また、剽窃の定義があいまいである、とも指摘している。論文の再利用は、そこに剽窃のあることを予想させるが、どの程度重複していれば剽窃ありと断定するのか。記者は30%以上の再利用に剽窃ありとしたらいかか、と提案している。また、同じく再利用であっても、二重投稿は剽窃とは違うかもしれない、と付け加えている。問題はこれらの調査が人手に依存していることにある。これを自動化できないか。

*

*

ここでギンスパーグが登場する。彼は、周知のように物理分野における最初のプレプリント・サーバArXivの開発者であった。

彼の学生ソロキーナは、そのサーバに投稿された28万論文を調査した。ここには重複した表現が30,316件、少なくとも677論文もあり、彼女はその22件を重複論文とし、うち3件を剽窃と判断した。

ソロキーナは学生剽窃発見用のアルゴリズムを応用した。このアルゴリズムは、新しいテキストをすでに刊行されたテキスト群のデータベースと比較するものであった。彼女は、少なくとも連続する6単語が2編の論文にある場合、あとの論文に「厄介なもの」ありと判断した。ただし、‘this work was supported in part by’ といった表現は省いている。

もう一つ、ソロキーナはテキスト群のデータベースをいかに用意したのか。学生の場合には公開されているWeb上の記事を収集し、それをデータベース化すればよい。彼女はそのようなデータベースとしてArXivを使った。

だが学術論文の場合、すべての先行論文がオープン化されているわけではない。先行論文を収集するためには「購読」という壁を越えなければならない。ソロキーナの場合、もし先行または後続の論文のどちらかが通常のジャーナルに投稿されていれば、それはArXivには含まれず、したがって剽窃を検出することはできない。

話を戻す。ギンスパーグはどんな意図をもっていったのか。重複が発見された場合、その事実を投稿者に伝え、あとの処置——撤回か投稿か——は投稿者に任せればよいと考えていたらしい。

*

*

剽窃疑惑をかけられた研究者は、またその論文を採用した編集者はどんな反応をするのだろうか。テキサス大学のロング他が、その調査結果を『サイエンス』（2009年3月6日号）に投稿している。彼らは、MEDLINEから、引用文献が類似しており、しかも著

者に重複のない論文を抽出し、Déjà vu（既視感）というデータベースを編集した。ここには212対の類似論文があった。

ここでは、後出論文が先行論文を引用する場合が22%、また、後出論文に含まれる図表が先行論文のそれに酷似するものが71%、さらに、後出論文が不正確な計算、矛盾したデータ値、複製または改変された写真を含む場合が43%、もあった。

ロングは、ここに163件の潜在的な剽窃ありと推定し、当の著者と編集者に質問状を送った。回答は144件であった。先行論文の著者の93%は自分の論文が剽窃されたことを知らなかった。

後出論文の著者は多様な反応を示した。28%は不正をきっぱりと否定した。35%は事実を認めて遺憾の意を表した。22%は自分は共著者にすぎず原稿の執筆に関係しなかったと釈明した。17%はなぜ自分の名前がそこにあるのか理解に苦しむと弁解した。

編集者はどのように対応したか。83のジャーナルは内部調査の組織を立ち上げ、46のジャーナルは当の論文を撤回処分にした。

だが問題が残った。当のジャーナルが撤回声明を載せても、それが自動的にMEDLINEに伝達され、MEDLINE上の該当情報に撤回のタグを付けるわけではない。なにも知らない読者はその事実を知ることができない。

さらに、インパクト・ファクターの問題がある。一般論としていえば、先行論文のほうが後続論文よりも高いインパクト・ファクターをもっている。だが、少数ではあるが先行論文と同程度の頻度で引用される後続論文もあった。ロングによれば、これはPubMedの検索アルゴリズムがより新しい論文を上位に示すというデフォルト・ルールをもっているためだという。つまり、剽窃が原著よりも目立ちやすいことになる。

*

*

話が行ったり来たりするが、剽窃発見のためには全ジャーナルを格納するデータベースが不可欠である。この点、ArXiveにせよMEDLINEにせよ不十分であった。どちらも、その研究分野における全ジャーナルをもっているわけではなかった。

だが、ここに網羅的なデータベースをもつ剽窃検出システムが開発された。それがCrossRefによるCrossCheckである（『情報管理』53巻3号, p.140-144）。ここには83の出版社が、48,517のジャーナルと書籍、2,550万件の記事を集積している。CrossRefは、領域横断的、出版社横断的な学術ジャーナルへのゲートウェイ・システムである。

CrossCheckは2008年に動き始めた。まず、テストが半年にわたり、3つのジャーナルに掲載された216論文ついてなされた。その結果、第1のジャーナルについては10%、第2のジャーナルについては6%、第3のジャーナルについては23%が剽窃であるとして受理を拒まれた。

CrossCheckに注目した『ネイチャー』（2010年7月8日号）は「剽窃を抑制する」という論説と「ジャーナルが剽窃対策を強化」というニュースを掲げ、CrossCheckに対する『ネイチャー』の意見を、主要なジャーナルの編集者の意見に重ねて示している。

CrossCheckは剽窃の発見に対して有効な手段ではあるが、その自動的な手段に頼りきってはいけない。人間による点検も欠くことはできない。例えば許される二重投稿がある。異なる分野のジャーナルに同一の論旨の記事を投稿することなど。また、英語の表現力の貧しい研究者が、序論において先行論文の一部を引用することもあるだろう。あるいは、総説に対しては「類似性すなわち剽窃」という判断について慎重であってもよい。

ところでCrossCheckの利用コストであるが、1論文あたり0.75ドルだという。