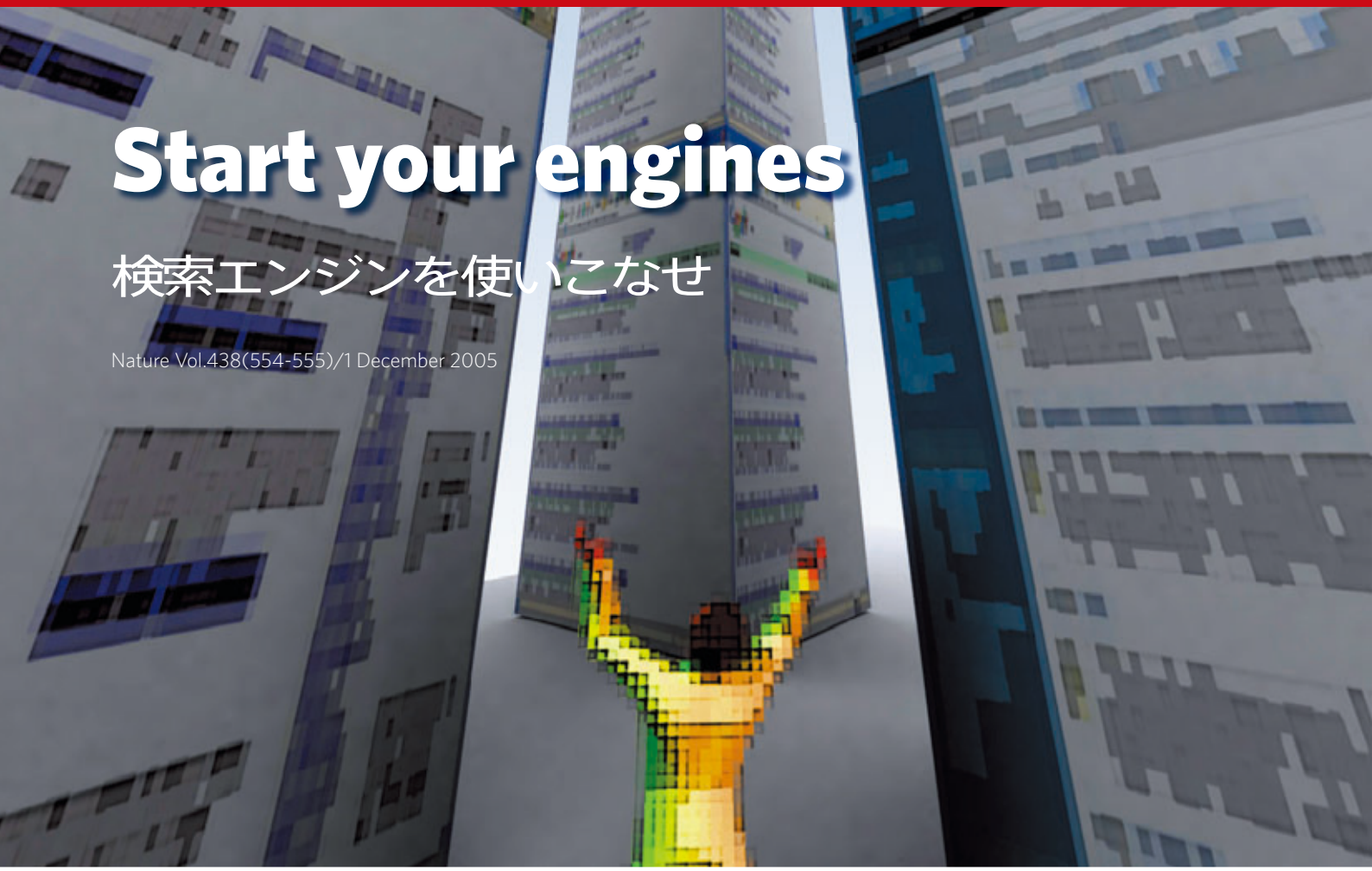

Start your engines

検索エンジンを使いこなせ

Nature Vol.438(554-555)/1 December 2005

C. DARKIN

Google 社が新たな検索サービスを立ち上げた。今回は科学者向けだ。この新しい検索エンジンは、利用者の望みどおりスムーズに動くのか。Jim Giles が検討した。

1980 年代半ば、インドの大学生だった Anurag Acharya は、欲しい論文を探し出せず、科学者に手紙を出さざるをえないことがあったという。穏やかな口ぶりで話すコンピュータ技術者の Acharya は、そのころを思い出して笑う。現在 Google 社で検索ツールを開発する彼は、それをインド人学生からイラン人教授にいたる、あらゆる人が一番に使うツールに仕立てたいと考えている。Acharya は「言語や分野を問わず、あらゆる学術情報を得るための定番にしたい」と話す。そして、そのような野心については、率直に「言うは易く、行うは難しだ」と認める。

単発の学術論文を探す人たちにとって、グーグル・スカラー（Google Scholar、以下スカラー）は理想的だ。無料でアクセスでき、使い勝手も本体の Google 検索エンジンと変わらない（コラム「検索の裏話」を参照）。しかし、すでに所属する大学機関などで文献情報源が購入されているような場合、そうした学者を、どうすればそれに乗り換えさせることができるだろうか。科学者の多くは論文を探すとき、米国立衛生研究所の PubMed や NASA の宇宙物理学データシステムをはじめとする特定分野のデータベースなど、使い慣れた手段を利用しているものだ。

2005 年 11 月の立ち上げ以来、Acharya のスカラーは評価が割れた。この検索エンジンの進歩を追跡するブログを立ち上げた図書館職員までいる。詳細な調査は行われていないが、教職員や学生がこの検索エンジンを使うようになってきたと話す図書館職員は多く、スカラーがほかのすでに確立されている高価な検索ツールに取ってかわるのではないかとみる者もいる。学術出版社のデータからも、スカラー利用者の急増がうかがわれる。*Nature* のウェブサイトも、学際的な科学検索エンジンとしてはスカラーからやってきたアクセスが最も多くなっている。

マックス・プランク神経生物学研究所（ドイツ・マーティンスリード）の神経科学者 Thomas Mrsic-Flogel はよく PubMed を利用していたが、スカラーを使うようになった。自分が何を探しているのかがよくわからないとき、この検索エンジンが役立つからだという。検索結果にはほかの論文への引用リンクが含まれているため、リンクをたどってい

くうちにおもしろそうなものに行き着くことがある。引用を追跡する機能がないPubMedでは、このようなことができない。Mrsic-Flogelは「引用のつながりを追いかけると、思いがけない論文にたどり着く。ほかではできないやり方で論文を見つけている」と語る。

この引用追跡のおかげで、スカラーは科学出版社が発売している従来の有料検索エンジンと直接競合することになった。エルゼビア社が検索エンジン「スコーパス（Scopus）」を発売した2004年まで、引用追跡はトムソンサイエンティフィック社の「ウェブ・オブ・サイエンス（Web of Science）」の独壇場だった。被引用数によって研究者や研究施設、学術誌は個々の論文の影響度を経時的に追跡することができ、多くの学者にとって悩みの種であり喜びでもある、インパクトファクターなどの指標が得られる。

スコーパスやウェブ・オブ・サイエンスと異なり、スカラーでは査読済み文献以外のものも検索できる。プレプリントアーカイブや会議議事録、また、無料でアクセスできる論文を置いておきがちな著者の個人サイトを保存していることの多い研究施設のリポジトリなど、従来とはちがう多数の検索ができるようになったことを評価する利用者もいる。こうした非公開文献（グレーリテラチャー）の重要性は増しているが、その明確な定義はまだない。一般的には、ある情報が別の学術情報に引用されていればGoogleは、その元の情報も学術的なものと判断していると考えられている。だが、オンライン出版が盛んになっていくなかで、こうした定義もまた変わっていくだろう。また、科学文献へのアクセス拡大を推進する立場からは、スカラーの登場で自分の論文を無料のオンラインリポジトリに置く科学者が増えることが期待されている。

しかし、スカラーは実際にどれほど機能するのだろうか。検索エンジンをいくつか使って体系的に検索を行った図書館職員は「使いものになる」と話す。100本を超える論文を対象とした2005年の調査では、対抗する商業検索エンジンと同程度の数の引用文献がヒットしたと結論づけられている[1]。ただし、その結果の解釈は慎重に行う必要があると情報科学者は話す。この調査では、引用文献のエントリーに重複や誤りが含まれていないかどうかが考慮されていないというのだ。

スカラーの検索結果をよくみると、重複がままあることがみてとれる。ハワイ大学（米・ホノルル）の情報科学者でスカラーに最も批判的なPéter Jascóは、何度もこの検索エンジンを試用した。そして、その結果をトムソンサイエンティフィック社のウェブサイトで冷評している。極端な例だが、「computers」（コンピュータ）と「intractability」（取り扱いにくさ）に関する文書の検索でヒットする上位100件には、*Computers and Intractability*という名の書籍に関する引用が92件（違いはわずかしかない）顔を出しており、この本と関係しないものは8件に過ぎなかったという。

ヒットしないことも

この問題の根源は、Googleが学術インデックスにレコードを加える際の方法にある。ウェブ・オブ・サイエンスとスコーパスでは、各ジャーナルの印刷版の文献抄録と参照文献をスタッフが調べ、出版社が提供する専用の電子媒体を利用する。対照的に、スカラーの処理は自動化されている。ソフトウェアロボットがウェブ上を巡回して科学論文とみられる文書を探し出し、続いて著者名や発行日などの関連情報を抽出するアルゴリズムが用いられる。これはきわめて安上がりで迅速な方法だが、今のところ更新が毎日行われているわけではなく、重複の削除や誤分類レコードの修正を行うための人の目によるチェックもない。

Googleが掌握する学術出版社の中には、論文の全文検索を認めているところが多数あるが、ウェブ・オブ・サイエンスなどでは概して抄録のみの検索に制限されている。しかし、スカラーのインデックスはオンラインソースに限定されているが、ウェブ・オブ・サイエンスのアーカイブは1900年にまでさかのぼる。また、自動処理ということでスカラーの引用追跡がおかしな結果を表示する場合もある。たとえば、ウェブ・オブ・サイエンスでポリメラーゼ連鎖反応に関する1988年の*Science*誌の論文[2]について調べると約14,000件の引用がヒットし、これが同誌史上最も引用された回数の多い論文であることがわかる。だが、スカラーでヒットするのは3,000件に満たない。

結局のところ、スカラーが探し出す非公開文献の引用と、元の文献から追跡される引用との間には、ほとんど共通部分がないと考えられる。たとえ被引用数がほぼ一致する場合でもだ。現時点で、図書館職員の考え方は一致している。徹底的な文献検索や正確な被引用数が必要ならウェブ・オブ・サイエンスやスコーパスを使い続けることだ。両エンジンは網羅する範囲が広く、レコードのインデックスが良好でエントリーの誤分類も少ない。図書館職員は、スカラーはまだ実験版、つまりベータ版にすぎないとも警告している。Google社は検索アルゴリズム、つまり何をインデックスとしているのかについての詳細を開示しようとしていないが、被引用数を確認するためのツールとしてスカラーを用いるのは時期尚早だという。

この3種類の検索エンジンは、いずれも進化し続けるだろう。スコーパスとウェブ・オブ・サイエンスは研究施設のリポジトリなど新たな情報源をデータベースに加える計画で、そうした情報源を検索するための新たな方法も導入しようとしている。たとえばスコーパスにはある化学データベースが組み込まれ、利用者は文献検索からそのまま目的とする分子構造の情報に行

検索の裏話

科学検索エンジンは文献検索に利用できるが、科学者は当然ながらさらに幅広い情報をウェブに求めている。本体のGoogle検索エンジンを利用した検索なら、多少手がかかる場合があるが、ちょっとした工夫で肝心な関連情報を早く見つけられるようになり、テーマについての多様な視点を得られる。

Google検索のアドバンスオプションでは、詳細な用語を使うことで検索を絞り込んだり、類義語で検索の幅を広げたりすることができる。ここでは薬のタミフル(Tamiflu)を例にとり、あまり知られていない検索の秘訣をいくつか紹介する。

Site:
ウェブサイトを探しあてるのに苦労することは多い。際限のないクリックに時間を浪費するのをやめるには、クエリーにウェブサイト名を入力してから「site:」と加えればよい。検索はドメイン名に限定することもできる。たとえば、「site:gov」では検索が米国政府のサイトに限定され、「site:nih.gov」では米国立衛生研究所に限定される。世界保健機関のサイトでタミフルを検索するときに「Tamiflu site:who.int」とすれば約100件がヒットする。検索の幅を広げて「tamiflu site:edu」などとすれば、米国内の大学の40,000件以上がヒットする。

Filetype:
スマートな検索に役立つのが、「filetype:」クエリーを利用した特定文書形式の検索だ。検索「Tamiflu filetype:ppt」では、会議の発表でよく用いられるパワーポイント形式のファイルのみが抽出される。「filetype:doc」とすればプロジェクトの企画書や政府文書が多くヒットし、「filetype:pdf」では科学的な情報が多く拾われるだろう。

Define:
この単純なクエリーでは、続けて入力した言葉の定義が各種のオンライン情報源から選び出される。たとえば「define:Tamiflu」というクエリーでは、何か国語かのウィキペディアで定義を見ることができる。

引用符
究極的にいえばウェブは人間に関するものであり、連絡先、または共同研究者を探すための方法はあるものだ。「"avian influenza" "workshop participants"」というクエリーではヒットが数百件現れ、上位のヒットには世界中の専門家の詳細な連絡先情報が得られるものが多い。どの科学分野でも同じように応用が可能だ。

Declan Butler

きつけるようになっている。しかし、この手の検索エンジンがスカラーほど手広く非公開文献を調べ上げるようになるとは考えにくい。エルゼビア社にはサイラス(Scirus)という別の無料検索エンジンがあり、ウェブ上の科学情報を検索できるが、引用文献を追跡することはできない。

> 「Google Scholarを、分野を問わずあらゆる学術情報を得るための定番にしたい」
> ― Anurag Acharya

では、Acharyaの大胆な目標はどうなるのだろうか。図書館職員は、あるテーマに関して通常2、3の重要論文が入手できれば十分という大学生や、不案内な分野についてさっとり早く調べたい研究者による検索が、スカラーの現在の高い利用率をもたらしているのだろうと話す。Acharyaはスカラーについて、そうした利用を目論んでいたと話す一方、専門分野の最新論文に常に触れておく必要がある学者を引き寄せたいともいう。トムソン社とエルゼビア社が新たなサービスへの投資を続行するなか、スカラーがそれに伍していくことができるのかどうか興味深いところだ。

ふたりの会社

Acharyaとともに働くフルタイムのスタッフは2名しかいないことから、Google社にとってスカラーは優先順位が低いのでは、と感じるかもしれない。しかし、同社がスカラーのデータベースを開発する過程で、社外のコンピュータープログラマーにソフトウェアを書いてもらうということも考えられる。それはかつて、同社が成果を出すために取った方法だ。スカラーでも同じ手法をとれば、図書館職員や学者たちが使い勝手や機能を拡張していってくれるだろう。

では、スカラーはソースを社外に公開するのだろうか。現時点でそれはない、とAcharyaはいう。ただ、多くを語ろうとはしないものの、その可能性は排除せず「構想に近づいてくれば再検討することになるかもしれない」と話す。

また、スカラーの盛り上がりが拡大し続ければ、同社のチームによる見直しもあるかもしれない。ヴァージニア工科大学(米・ブラックスバーグ)の図書館職員はすでに、インターネットブラウザ用にLibXというフリーの拡張ソフトウェアを開発した。これを利用すると、強調表示されているテキストをマウスでクリックするだけで、スカラーから論文を引き出すことができる。LibXを使えば、どんな文献もそこにありさえすれば、コンピュータ上から直接アクセスできる。そして、そういったものこそがGoogle社、図書館職員双方が意とするツールなのである。 ■

Jim Gilesは*Nature*のシニアレポーター(ロンドン)。

1. Bauer, K. & Bakkalbasi, N. *D-Lib Magazine* 10.1045/september2005-bauer (2005).
2. Saiki, R. K. *et al. Science* **239**, 487-491 (1988).