第 2 回 SPARC Japan セミナー2011

「今時の文献管理ツール」ワークショップ

文献管理と文献推薦を一体化する生命科学研究者のための統合文献ソリューション

岩崎 渉

（東京大学大気海洋研究所 地球表層圏変動研究センター 講師）

講演要旨

21 世紀に入り、ゲノム、オーミクス、バイオインフォマティクス（生命情報科学）といったキーワードに代表される技術革新が生命科学に大量データをもたらした。これまではデータを産出することが何よりも重要であった。今日、データはすでに大量に存在し、それを解釈する段階がボトルネックとなりつつある。大量データを解釈するためには、専門分野に限らない幅広い分野の最新動向を把握し、かつ、そうして得た知識を死蔵せずに活用していくことが不可欠となる。我々はこれらの問題を解決するため、文献管理と文献推薦を一体化する統合文献ソリューション TogoDoc を開発した。指定したフォルダに文献 PDF ファイルを保存するだけで自動的に文献情報の解析が行われ、ユーザの研究分野や好みが解析され、文献データベースに毎週 10,000 件以上登録される膨大な新規論文の中から必読論文が推薦される。Windows 版と MacOSX 版が http://tdc.cb.k.u-tokyo.ac.jp/ からフリーでダウンロード可能である。

岩崎 渉

専門分野はバイオインフォマティクス。2005 年東京大学理学部生物化学科卒業、2009 年東京大学大学院新領域創成科学研究科情報生命科学専攻博士後期課程修了（博士(科学)）。同専攻助教を経て 2011 年より現職。情報・システム研究機構ライフサイエンス統合データベースセンター業務協力者、日本バイオインフォマティクス学会評議員・幹事。

本日はわれわれが作っているツールである TogoDoc についてご紹介させていただきます。われわれは、分野を絞り込むことで、より高精度な文献の推薦と文献の管理を一体的に提供することができるのではないかと考え、このプロジェクトを始めました。Mendeley にもこれから推薦機能が追加され、10 月に発表された ReadCube も文献管理と文献推薦をリンクして提供するということです。このような推薦と管理を一体的に提供するというアプローチがこれから主流になっていくのではないかと考えています。

論文大量出版時代における効率的な知識の抽出

現在、大量の文献が毎日出版されており、医学・生物学分野の文献だけで年間およそ 80 万本の論文が出ています（図 1）。1 年を分数に直すとせいぜい五十数万分ですから、1 分に 1 本以上のペースで論文が出版され続けていることになります。特に医学・生物学分野では近年、ゲノムあるいはトランスクリプトームな

ど、オーミクス研究において一度に大量のデータが得られるようになっています。このようなビッグデータのサイエンスの時代には、自分の専門分野の専門誌だけを読んでいればよいのではなく、複数のさまざまな分野の知識を素早くアップデートしなければいけません。すなわち、たくさん出版される論文の中から、より効率的に有用な知識を抽出していくことが大切になります。また、すでにご存じのようにもう一つのバックグラウンドとして、今日、研究者のコンピュータの中に大量の PDF 論文が存在しているということがあります。

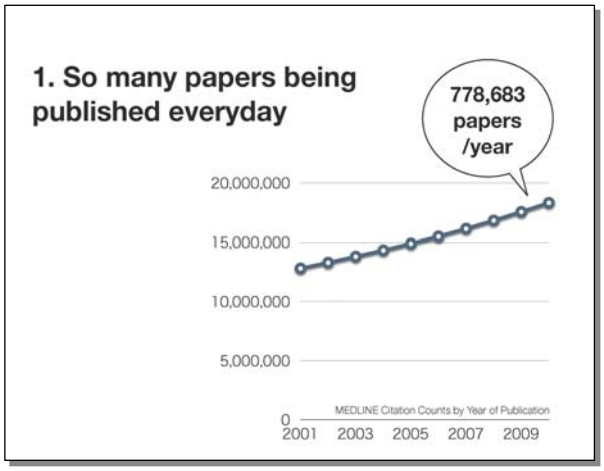
この二つの点を結んで一つのソリューションを統合的に提供するのが TogoDoc です。

TogoDoc の概要

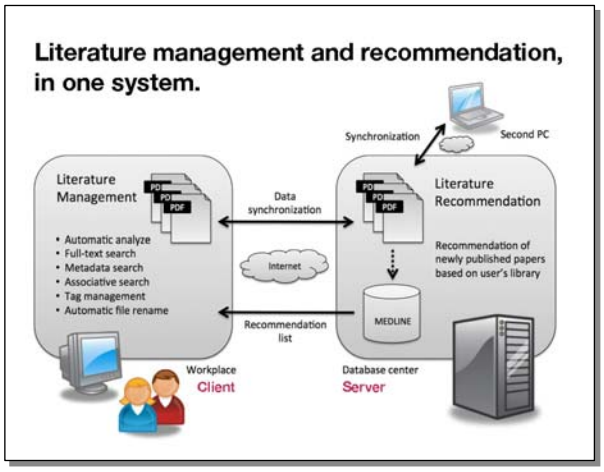
TogoDoc は Windows と MacOSX で動作します。英語版と日本語版があり、無料でダウンロードできます。

TogoDoc の開発プロジェクトは 2007 年 8 月に始まりました。その後、2008 年に Alpha Version と Beta Version、2009 年に完全版をリリースしています。東京大学と情報システム研究機構のライフサイエンス統合データベースセンターにおいて開発が行われています。開発をメーンに行っているのが日本人であるということは、ほかのツールと決定的に違う点の一つでしょう。これは些細な問題かもしれませんが、日本のユーザーや図書館の特殊な事情に対応できる可能性があると言えます。

われわれのツールの全体像です（図 2）。今はどんなツールもそうですが、デスクトップで動くツールと、サーバー上で動くツールがインターネットを介して連携しています（クラウド）。研究者がそれぞれのコンピュータで使う文献管理機能に関しては、自動的に PDF ファイルを解析する他、フルテキストやメタデータ（タイトルや雑誌名）による検索が可能となっています。さらに、タグを付けたり、ジャーナルのサイトから PDF をダウンロードした際にばらばらなファイル名になったものを自動的に変えたりする機能が付いています。



（図 1）So many papers being published everyday



（図 2)Literature management and recommendation, in one system

ます。

PDF ファイルについては、データを自動的にサーバーと同期し、7GB まで無償でアップロード可能になっています。これはほかのツールやサービスに比べて比較的多いサイズになっています。7GB の文献を持っている方はめったにいないと思うので、ほぼ無制限に使えると言ってよいでしょう。

そして、アップロードされた PDF ファイル群から研究者の好みを読み取り、それに応じて必読文献を自動的に推薦するのが、TogoDoc の推薦機能です。このように文献管理と推薦を一体的に提供するシステムとなっています。

TogoDoc のインターフェース

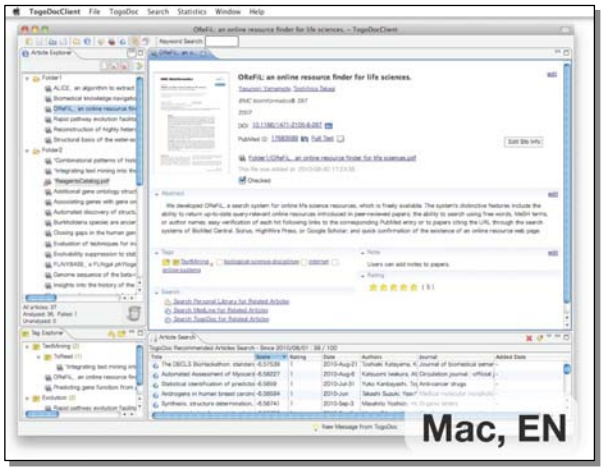
Windows 版（日本語版）と、Mac 版（英語版）のスクリーンショットを使って、インターフェースについてご説明します（図 3、4）。

インターフェースには四つのウインドウ（ペイン）があります。左上の Article Explorer（論文エクスプローラー）は、単純にユーザーが持っているフォルダーをファイルマネジャーとして表示するものです。右上の Literature Tab（文献タブ）は、今選択されている文献の詳細な情報を表示します。左下の Tag Explorer は、文献にタグを付けることができ、また、タグごとに文献を探すためのウインドウです。右下の Search/Recommendation Pane が大事なペインで、ユーザーが TogoDoc を立ち上げるごとに、新しく出版された論文の中から読むべき論文がここに通知されます。また、それぞれのペインはドラッグ＆ドロップで自由に大きさを変えられます。

少しだけ実際に動かしてみましょう。われわれは、論文を探すときに 1 ページ目の画像が非常に大事だと思っています。これをトップに置き、視覚的に見付けやすいようにしています。タグを自由に付けられることはもちろん、適切なタグを付けるのがなかなか大変だという意見も伺いますので、タグの自動サジェスチョン機能も付けています。論文に対するノートを書く欄もあり、評価やレーティングを記入することができます。当然、フルテキストサーチなどの検索機能があります。

他に面白い機能としては、タグやフォルダーを選択すると、そこに入っている論文の著者を一覧形式で出すことができるので、例えば、その分野のシンポジウムの講師あるいは論文のレビュアーとして適切な方を探すのに役立ちます。

（図 3）インターフェースについての説明：Windows 版（日本語版）



（図 4）インターフェースについての説明：Mac 版（英語版）

推薦機能

右下に出ているのが、PubMed レコメンデーションという、どんな文献を持っているかを解析して、その研究者が興味を持ちそうな文献を自動的に推薦する機能です。この推薦には、Mendeley の場合は協調フィルタリングを使っているのではないかと思いますが、われわれは論文の中身を見て推薦しています。例えば、医学・生物学分野では今、学際的な研究が非常に多くなっていますが、多くの研究者が読んでいるから読むという論文だけではなく、自分のニッチに合う論文を読みたい場合があります。ですから、アマゾンのように、ほかの人が買っているからこれも買いなさいというタイプではなく、例えば分野 A と分野 B の論文が多い研究者には、その 2 分野の境界分野の論文を推薦します。

文献の中身を見て推薦する場合に大事なのは、内容は同じでも違う言葉で表されるという問題です。これ

を解消するために、医学・生物学分野のためにつくられたボキャブラリーのリスト（オントロジー、タキソノミー）、MeSH terms などを使って、同じ分野を違う言葉で示している場合にもきちんと推薦が行われるようにしています。

それから、当然、ユーザーである研究者がよく読む論文著者やジャーナルといった情報も取得することが可能なので、それらの情報も用いています。これらの様々なパラメーターを医学・生物学分野用にチューニングすることで、本当に読みたい論文が出てくるようにしています。このようなことは、やはり分野を絞るとしやすくなると思います。

その他の便利な機能

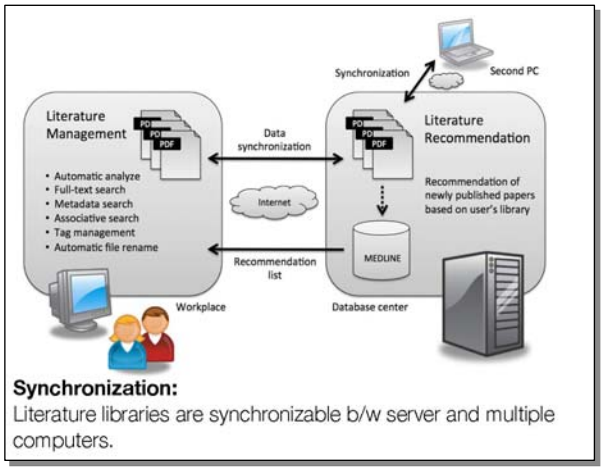
その他、PDF の自動的な名前変更の機能もあります。それから、データをサーバー上に同期し、さらに別のコンピュータとも自動的に同期することが可能です（図 5）。今の研究者は複数台のコンピュータを持っているのが普通なので、このようになっています。当然、サーバーにウェブブラウザでアクセスすることで、どんなコンピュータからも自分のライブラリにアクセスすることもできます。

ウェブブラウザからアクセスすると、さらに幾つかの機能が使えます。推薦に使われているキーワードの重要度を自分でチューニングしたり、あるいは毎回 TogoDoc で見るのではなく、iPhone などのスマートフォンから必読論文をチェックすることもできます。

これは使っているテクノロジーです（図 6）。TogoDoc にご興味をお持ちの方は、ぜひインターネットで検索していただければと思います。使用している具体的なテクノロジーについても論文が出ていますので、そちらもご覧ください（図 7）。

大量情報時代の研究を支援する TogoDoc

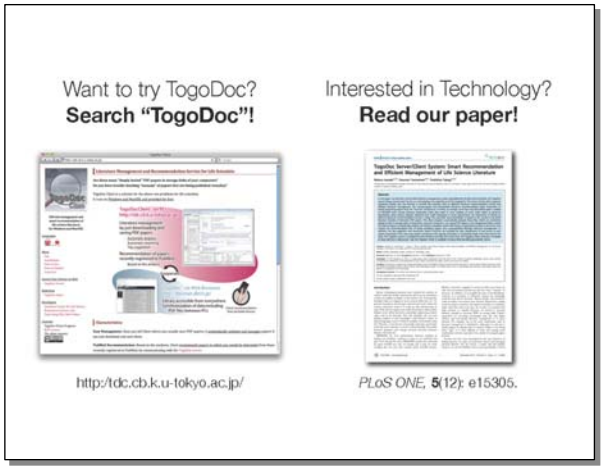
現在、情報が非常に過多になっています。われわれの分野である医学・生物学分野で言うと、毎日新しいバイオインフォマティクスのツールやデータベースが



（図 5）その他の便利な機能：Synchronization

Technology Keywords
Java: TogoDoc software is implemented in Java.
Eclipse Rich Client Platform: TogoDoc software platform.
JPedal: Java library for analyzing PDF.
LAMP: Server is deployed using Linux, Apache, MySQL, Perl.
Tokyo Cabinet: Server-side database manager.
OpenID: Authentication protocol for sharing a single account.
JSON: Data transfer format b/w server and client.
RSS/ATOM: Data transfer format for RSS readers.
PubMed/MEDLINE: Literature databases at NCBI.
NCBI E-Utilities: API to access PubMed.
Lemur/Indri: Information retrieval toolkit.
MeSH: Subject headings in MEDLINE.
BibGlimpse: Bibliographic analyzer.

（図 6）Technology Keywords



（図 7）Search “TogoDoc”！/ Read our paper!

出てくる状況になっています。しかし、実際には研究者は古いツールやデータベースを使っていて、新しいツールが出たことに気が付かないということが日常茶飯事のように起こっています。それから、もはや文献だけではなく、研究のスピードが速くなっているので、メーリングリストやニュース、ブログ記事、SNS のポストなどから情報を得ていかなければなりませんが、情報量が多すぎてどうしても追い付いていけません。例えば私の場合、1 日に 50 件くらいメールが来るのですが、その中で私が本当に興味を持つメールはごく一部です。ですから、やはりこれからの時代には、たくさんの情報をフィルタリングするための技術が必要であろう、あるいは新しいデータベースをサジェストしたりする機能が重要であろうと私たちとしては考えています。

このような時代に有用なのは、文献のセットを、研究者がどのような興味を持っているかを表すためのツールとして使うことです。文献のセットは非常にコンパクトで、しかもコンピュータ上で扱いやすい形式のデータになっています。これを使って研究者に、論文に限らず、さまざまなテキストやメッセージについても、「これはあなたにとって大事です」というようにサジェストすることが将来的には可能になるのではないかと考えています。

開発を通して学んだこと

繰り返すまでもありませんが、文献は、その時代に研究を進める上で非常に重要な鍵となります。しかし文献の出版には、法的な問題、エコノミカルな問題、レビュアーにかかる負担、社会的問題等、さまざまな問題があります。それらを解決すると同時に、文献をうまく扱うための情報技術を研究・開発するという方向性が非常に重要になるという問題意識でわれわれはやってきました。ただ、私自身は本職はバイオインフォマティクスの研究で、DBCLS の研究者もそれぞれ研究テーマを持っています。別の研究テーマを持ちつつ、このようなこともしたいとなると、スピード感が損なわれ、やはり急激な開発はどうしても望めません。われわれは、これは研究者のためのプロジェクトなので研究者が率先して取り組むべきだと考えていたのですが、恐らくアカデミアだけではなく、ビジネスあるいはコマーシャルのセクターとも協働してゆくべきではないかと今、思っているところです。そのような方法についてのご見識があれば、ぜひ伺えればと思います。

●Henning　文書管理と、情報の洪水の中でいかにこれをうまく使っていくかという点で、研究者の役に立つものをつくりたいというビジョンはわれわれと非常に共通していると思います。推薦（レコメンデーション）のアルゴリズムについて、少し私の方から情報提供をしたいと思います。私ども Mendeley の推薦機能は、アマゾンのようなコラボレーション型ですが、それだけではなく、MeSH terms のキーワードも見ています。それから、ドキュメントにユーザーが付加したキーワードもフィルタリングに使っています。文章に関連性があるかどうかを知る上で、キーワードは予測子として非常に有用です。ですから、API を通して Mendeley と TogoDoc が協働できる可能性があるのではないかと思います。私ども Mendeley の研究開発チームと連絡が取れるようにしますので、技術についての具体的な情報を得ていただければと思います。

●**岩崎**　ありがとうございます。Mendeley でどのように推薦をしているのかについて、残念ながら情報が見付かりませんでしたので、今お話しいただいてよかったです。喜んでコラボレーションしていきたいと思います。