SPRINGER NATURE

springernature.com

テキストデータマイニング白書

ADVANCING DISCOVERY

データに見識をもたらす：
テキストデータマイニングにおけるインフォプロの役割

– Mary Ellen Bates

情報プロフェッショナル（インフォプロ）、知識労働者、図書館員は、膨大な量の情報管理と検索に長年携わっており、中には Springer Nature などが提供するオンラインサービスの購読を管理したりその付加価値を評価したりする人々もいます。彼らは、研究者のために専門的なデータセットを特定して取得するほか、検索可能な内部リソースやコレクションを作成・管理します。インフォプロが研究情勢を把握する方法は、膨大なデータセットにテキストデータマイニング（TDM）を行う、洗練されたツールが開発されたことで大きく進歩しました。インフォプロは、組織内で情報がどのように使用されるのかを把握すると同時に、どうすれば情報を検索しやすくし、その価値を高められるかについても熟知しており、TDM プロジェクトに独自の展望をもたらします。

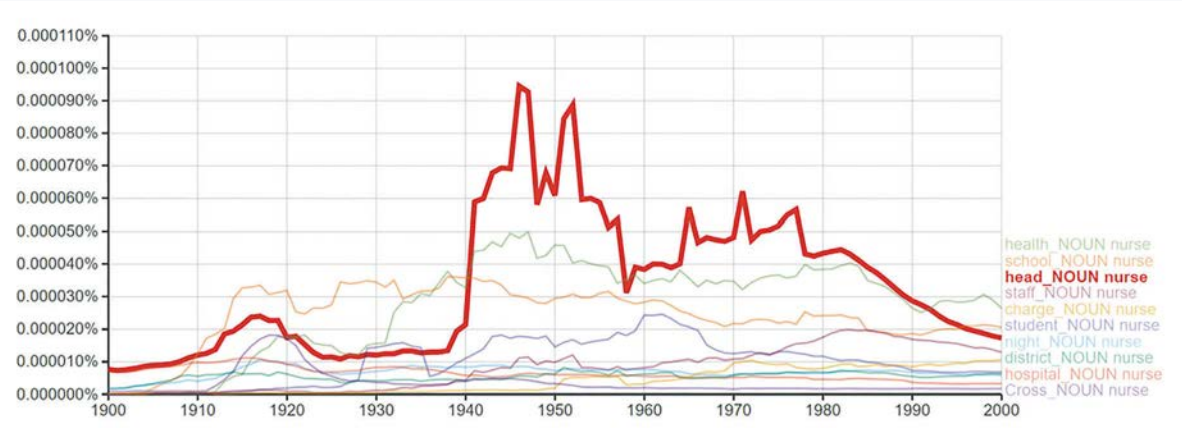
TDM の目的は、情報のフィルタリング、データの特定、データ間の関連性やパターンの発見です。革新的なのは、研究者がどのような具体的な質問をすべきか分からなくても、データセットを探索できるという点です。

TDMの基本

TDM プロジェクトでは通常、引用に関する書誌データベース、研究論文のアブストラクトや権威あるレファレンスソースなど、巨大なデータの集合体から始まります。それぞれのレコードが分析された後、個別の情報が TDM ツールにより、構造化された形式で抽出されます。このような個別の情報単位は「セマンティックトリプル」と呼ばれ、知識の断片を反映させた 3 つの要素で構成されており、主語 – 述語 – 目的語の形式で表現されます。たとえば「*The sky is blue*」のファクトは、「*the_sky – has_the_color – blue*」のようなトリプルで表現することができます。同様に、書誌レコードから生成されたセマンティックトリプルには、「*this_article – has_the_author – John_Doe*」や「*John_Doe – is_affiliated_with – Drexel_University*」などが含まれます。書籍のチャプター、組織のプロファイル、著者といったあらゆるタイプの情報にデータセットを適用することで、非常に簡単に変換できるようになります。

Google Books Ngram Viewer（books.google.com/ngrams）は、書籍のフルテキストに活用された際、TDM がいかに有用かということを証明しています。この Google プロジェクトでは、数百万の書籍のデジタルコンテンツを分析し、それぞれの単語と文章を解析しています。たとえば、文章中の単語は、意味と関連性で分析されます。「school」という単語は名詞「nurse」を修飾する形容詞であり、また、主語「nurse」は目的語「boy」に「treated」という行為を行います。我々は、Ngram Viewer を照会することで、「nurse」という単語が（「nursing someone back to health（元気を取り戻すまで誰かを看護する）」のような動詞ではなく）名詞として使用されている事例を探すことができます。クエリにより、あらゆる形容詞を検索してソートし、もっとも頻繁に使用される上位 10 のフレーズや、時代を通じてのその相対使用頻度を識別することもできます。図 1 では、「head nurse」というフレーズの使用頻度が 1940 年代と 1950 年代にピークを迎えていることを示しています。

図1　Google Books Ngram Viewerの検索結果



Google Books Ngram Viewer は、書籍に対し見識を与えてくれますが、TDM プロジェクトでは複数のデータベースやデータコレクションを含めることができます。たとえばインフォプロが、クエリを使用する API によって研究者の発見プロセスを改善したい場合や、医学概念に言及する検索語を特定し、それぞれの概念を米国国立医学図書館の Medical Subject Headings（MeSH）の階層構造シソーラスで見つけたい場合、オリジナルのクエリを拡張することで、その医学概念の MeSH 記述子だけでなく、その概念の範囲に入るあらゆる記述子を含めることができます。たとえば「*Opium Dependence*（オピオイド依存）」を検索する場合、MeSH 記述子、Opioid-Related Disorders [C25.775.675]、Heroin Dependence [C25.775.675.400]、Morphine Dependence [C25.775.675.600]、Opium Dependence [C25.775.675.800] などを含めるよう拡張します。

現在、インフォプロはあらゆるデータベースに含まれるデジタル情報だけでなく、さらに多くの情報にさらされているという問題に直面しています。彼らの関心事は、どうすればそれぞれの情報の中に埋もれた知識を明らかにできるのかということです。結論から言うと、論文やレポートの全文にアクセスできるだけでは、必ずしもコンテンツの最上の形式とは言えません。現在では、個々のレコードに含まれる情報だけでなく、こうした断片的な情報との関連性が重要になっています。

ある大手製薬会社の情報科学者は、TDM によりアブストラクトだけでなく全文をベースにしながら情報の関連性を見つける方法を紹介しており、TDM の価値はフルテキストの可用性とともに増大すると述べています。「通常の検索では限界がある場合でも、TDM はオントロジーを通じ、非常に多くの同義語を用いて複雑な検索を行うことができます。さらに、具体的な質問に対する関連部分や情報を膨大なテキスト文献や特許情報から抽出することもできます。単にヒットした文書のリストが表示されるだけではありません。」

「通常の検索では限界がある場合でも、TDM はオントロジーを通じ、非常に多くの同義語を用いて複雑な検索を行うことができます。さらに、具体的な質問に対する関連部分や情報を膨大なテキスト文献や特許情報から抽出することもできます。単にヒットした文書のリストが表示されるだけではありません。」
ある大手製薬会社の情報科学者

APIによるコンテンツマイニング

インフォプロは、情報へのアクセスを容易かつ効率的に行えるよう、長年APIを用いてきました。また、オープンアクセスコンテンツとスマートデータアノテーションの拡大に伴い、情報発見能力が向上し、さらにオプションが増えています。たとえばSpringer Natureは、インフォプロや研究者がSpringer Natureとそのパートナーである学術機関双方のSTMコンテンツに対し、新たな洞察を得られるAPIを作成しました（詳細については**dev.springernature.com**を参照のこと）。オープンデータにリンクしたAPIの簡単な利用例としては、ある論文に対してDOIに基づいた引用数を作成することがあげられます。これは、検索したコンテンツに目を通す研究者だけでなく、所属組織の研究の影響度を評価する意志決定者にとっても価値ある情報となります。また、Springer Nature Journal Suggester（**journalsuggester.springer.com**）は、研究者が投稿に最適なジャーナルを選定する際に役立ちます。著者がタイトルとアブストラクトを指定すると、APIがそれぞれのタイトルにインパクトファクターと採択率を付与した、推奨ジャーナルのランキングリストを返します。Springer Natureではコンテンツに化学化合物アノテーションを追加しており、これを活用することで研究者のクエリによりAPIが検索を拡大して、その化合物の同義語を持つコンテンツを検索できます。

競争の激しいインフォプロもまた、隠れた情報を検索すべくAPIを活用しています。分野の競合他社のニュースや新規参入企業の公開済みの書類や会議録のほか、隣接市場の進展についてもモニタリングできます。またAPIは、Glassdoor.comなどの求人サイトからも見識を得られるため、競合他社がどのようなトップスキルや専門職の人間を採用しているか特定できます。

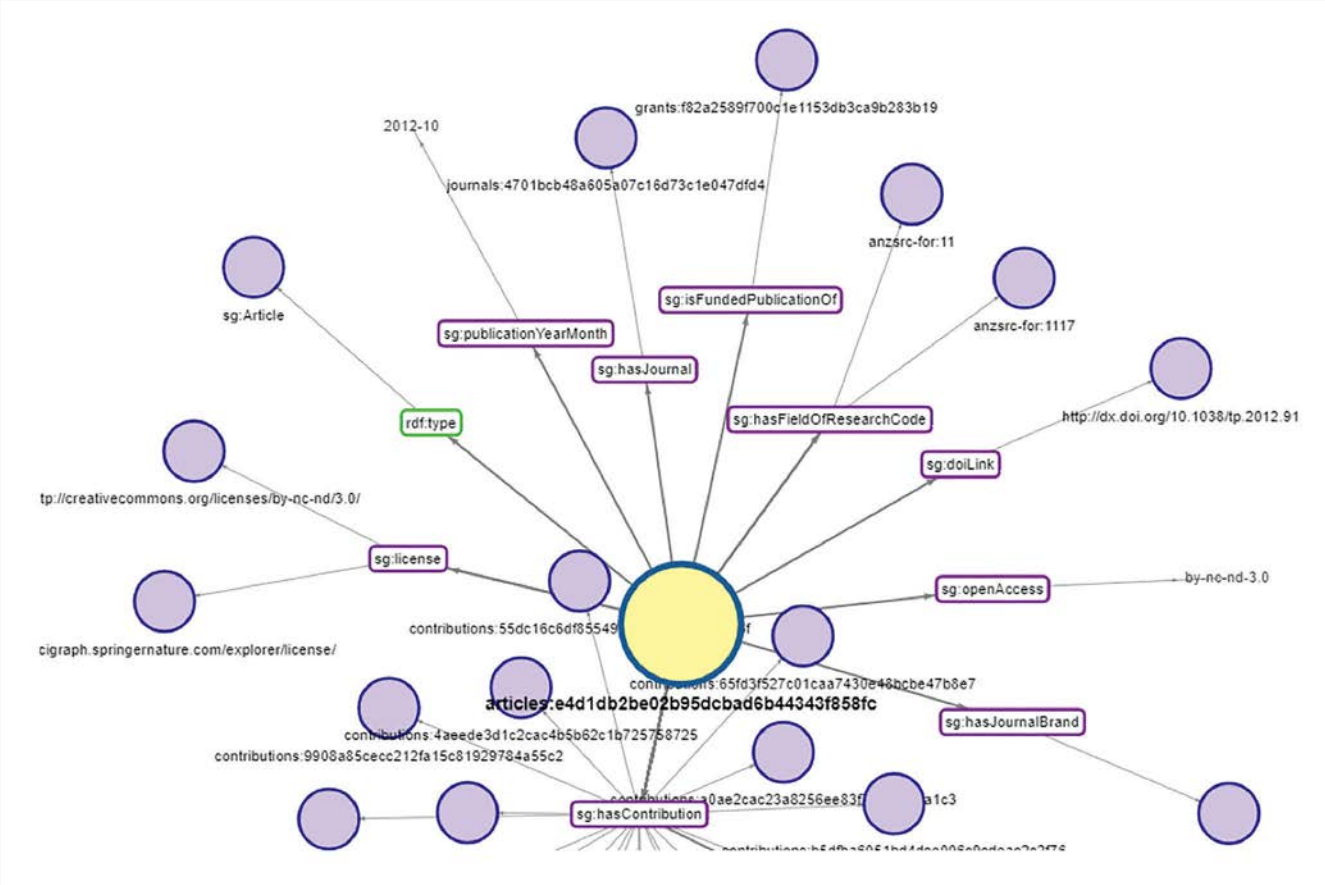
Springer Nature SciGraph

インフォプロの多くが関心のある論文や書籍のチャプターの検索、引用リストやドキュメントのコレクションという想定されるアウトプットからSpringer Natureに馴染みがあることでしょう。新たなオープンデータリンクのTDMプラットフォームであるSpringer Nature SciGraph（**springernature.com/scigraph**）では、大きなデータセットをビジュアルパターンで表現できるため、書誌検索単独では答えが出ない質問にも対応することができます。Springer Nature SciGraphでは、TDMツールを用いてジャーナルと論文、書籍とチャプター、組織、機関、資金提供者、研究助成金などのメタデータを組み合わせ、学術情報がどのように関連しているのか、分野、フォーマット、コンテンツタイプを超え確認することができます。

優れたメタデータは、信頼性の高い優れた見識をもたらします。Springer Natureは、文献内の化学物質、化合物、分子への言及について認識やアノテーションを行うソフトウェア企業、InfoChemと提携し、既存の文書で物質がどのように言及されているかに関係なく、関連物質の検索に一貫性を保つようにしています。これにより、情報の発見可能性の向上だけでなく、研究情勢の理解にも大いに役立つことでしょう。

たとえば、*Translational Psychiatry*の2012年10月号に掲載された「Adherence to a Mediterranean diet and Alzheimer's disease risk in an Australian population（オーストラリア人の地中海食への固執とアルツハイマー病のリスク）」という論文を取り上げてみましょう。図2は、この論文に対するSpringer Nature SciGraphのレコードを示しており、Springer Nature SciGraph Data Explorer（**https://scigraph.springernature.com/explorer**）で表示したものです。論文へのいくつかの関連性がハイライトされています。

図2　Springer Nature SciGraph Data Explorer



円の間のそれぞれのリンクは、セマンティックトリプルを示しています。上記のトリプルには、次のような関連性が含まれています。

- Article:e4d1db2… hasSubject subjects:alzheimers-disease
  （この論文は、アルツハイマー病に関するもの）
- Article:e4d1db2… hasFieldOfResearchCode anzsrc-for:1117
  （この論文は、1117 のコードが付与された「public health and health services
  （公衆衛生および公共医療サービス）」の研究分野に関するもの）
- Article:e4d1db2… hasContribution contributions:0934d8…
  （この論文の寄稿者の 1 人は S L Mathieson、個人コードは 0934d8…）
- Article:e4d1db2… hasContributingOrganization grid-institutes:grid.1008.9
  （寄稿者の 1 人は、grid.1008.9 のコードが付いたメルボルン大学に所属）
- Article:e4d1db2… isFundedPublicationOf grants:f82a25…
  （この論文は、grants:f82a25…のコードが付いた『Mediterranean Diet and Other Dietary Patterns in Alzheimer's Disease』というタイトルで助成金により資金提供を受けている）

この論文に対するその他のトリプルには、書誌サイテーションの全ての要素（タイトル、ジャーナル名、日付、巻号、DOI、アブストラクトなど）、著者の所属先、アブストラクト、言語などが含まれています。Springer Nature SciGraph のその他のレコードでは同様に、著者の所属機関、具体的なプロジェクト名と助成金、非営利組織などについての詳細情報が提供されています。

インフォプロの役割とスキル

効果的な TDM プロジェクトは、非常に有能で多くのつながりを持つ研究者のようなものです。こうした研究者が、それぞれの新規プロジェクトに何をもたらすか想像してみましょう。

- 専門ジャーナル、会議録、書籍、映像、オンライン・セミナー、レポート、特許など、専攻分野に関する文献を定期的にモニタリング。
- 学術会議に出席し、関連分野の研究者と顔を合わせ、最新のプロジェクトに関する情報を入手。
- 同僚と協力し、査読済みジャーナルに研究結果を発表。
- 専攻分野の関連組織や資金提供者をモニタリングし、助成金のパターンを分析。

この研究者は、様々な情報源から得た情報に精通しているため、通常では見落としがちな概念の傾向やその関連性について把握することができます。たとえば彼女が以前聴講した学会発表や、大学への助成金の近年の増加傾向をもとに、韓国での新たな進展に注視すべきことを知っていたとします。この研究者は、自身が追うあらゆるトピックに自分なりの分類法があるため、関連する概念の関係性に直感的に気付くことができるのです。

専門分野で利用可能な全ての情報を含めることで、その研究者の展望が飛躍的に拡大するという状況を想像してみてください。また、想像できるあらゆる研究分野に同様の超人的な研究者が現れたらと想像してみてください。TDM が組織に提供できるのはまさにこうしたものであり、情報や知識管理プロジェクト同様、インフォプロが重要な役割を果たすことができるでしょう。

研究者がさらに多くのデータ分析スキルを持ち込み、より多くの無料または購読サービスによる情報が利用可能になるにつれ、特に TDM において情報を検出、強化、管理、維持する方法に熟知したインフォプロの必要性が増していきます。

TDM の価値は、どのソースを含めるか、どのような関連性をモニタリングするか、特定のプロジェクトにおいてどのようなタイプのメタデータが必要かをくみ取ることで決まります。インフォプロが正しい質問を設定できれば、より大きなコンテクストが判明するだけでなく、優れた見識をもたらす具体的な情報セットの特定にもつながります。インフォプロは、限界、制約、各ソースのコストを勘案しながら、どのリソースを使用すべきか把握しています。また、研究者が情報をどのように活用するか、つまり彼らの問題へのアプローチ方法、情報探索行動のほか、次に得た情報で何を行うかについても理解しています。加えて、クライアントがどのような専門検索語を使用すべきか、どのように複数の情報源のデータを結合させるのかといったことには関心がないことを承知しており、クライアントが質問から洞察へとできる限り速やかに移行できるようなポータルや API を構築します。

インフォプロは、行政機関のデータセットやオープンデータ・イニシアチブ、Dryad（datadryad.org）や ICPSR（icpsr.umich.edu/icpsrweb/index.jsp）といった既存の科学出版物の集合的データリポジトリ、Springer Nature のような商業サービスなど、どのデータソースに注目すべきか熟知しています。

過小評価されているインフォプロのスキルに、いわゆる「レファレンスインタビュー」、より正しくは情報ニーズインタビューと呼ばれるものがあります。インフォプロは、研究者に情報から価値ある見識を見つけるための正しい TDM ツール、正しいデータセット、正しいアプローチを結びつける前に、研究者が思いもよらない点を含む、潜在ニーズを理解しなければなりません。インフォプロはまた、明快な答えのない問題への対応にも習熟しています。クライアントである研究者にとっての問題解決とは、様々なソースから情報を引き寄せ、他のグループとの協同により、どのようにして解答にたどり着くかを明らかにすることだと知っ

ているのです。

Content Capital LLC のオーナーで、TDM の業界オブザーバーでもある Scott Attenborough は、次のように述べています。「確かにインフォプロは、正しいクエリを作成し、階層を構築するスキルを持っていますが、本当に面白いのは、担当するクライアントのビジネスについて学ぶことなのです。クライアントは往々にしてどんな質問をすれば良いのかも分かりません。そのため我々の仕事は、それぞれのクライアントの使用事例を理解した上で、誰がどの分子の研究をしているか、この企業がその疾患にどのように取り組んでいるかといった、クライアントにとって何が重要か理解するのに役立つ正しいツールを作成することなのです。」

しかしながら、インフォプロが様々な情報ソースに精通していることが却って妨げとなることがあります。彼らは書誌データの検索、数百万の論文の念入りな調査、管理可能な論文数を研究者が検索できるまでさらなる絞り込み検索を行うことに慣れています。これに対し、TDM プロジェクトでは、パターンの検索や情報ピースがどのように組み合わっているのかを探りながら、想定外の見識を調べます。検索者は、リサーチを始める段階では、何を見つけられるのか分かっているわけではありません。「解答」は論文の集合になることもあれば、一連の図形になることもあります。

オンライン検索者に共通する問題は、一貫性のあるインデックスが付与されていないトピックに関連する情報の発見が難しい点です。医薬品は、執筆者の出身国、言語や習慣に基づき、別々に言及されることがある上、事項索引に全ての構成要素が含まれているとは限りません。たとえば米国では筋萎縮性側索硬化症またはルー・ゲーリッグ病と呼ばれている疾患が、英国では運動ニューロン疾患と呼ばれているといった、疾患が異なる名称で呼ばれている場合です。また、同じ言葉が文脈によって異なる意味を有することもあります。「*hearing aids*（補聴器）」と「AIDS（後天性免疫不全症候群）」などがこれに該当します。特定のガン、たとえば網膜芽細胞腫や多発性骨髄腫についての論文では、ガン（*cancer*）という言葉に言及しないこともあります。複数のデータセットを同時に検索する場合、用語に一貫性のないことで問題がより大きくなります。

TDM プロジェクトでは、権威あるデータセットを追加することで、この問題に真っ向から取り組むことができます。すなわち、あらゆるバージョンの概念を単一の権威あるエントリにリンクさせることでまったく異なる情報を検出可能にします。DBpedia を例に挙げます。DBpedia（**wiki.dbpedia.org**）はクラウドソースのオープンデータプロジェクトであり、信頼できる情報に基づきセマンティックナレッジグラフを作成します。Wikipedia の構造化データを抽出し、一貫性のある検索可能なフォーマットによりその情報のデータセットを作成します。Springer Nature、Eurostat、BBC といったコンテンツプロバイダは、DBpedia のバックリンクをそのコンテンツに統合することができ、コンテンツの検出可能性を向上させ、研究者はデータから新たな見識を明らかにすることができます。

ある大手製薬会社の情報科学者は次のように述べています。「TDM により、信頼性と正確性の高い数値 / 定量的情報（投与量など）が見つかり、メタデータとしても抽出できます。これはその他のパラメータの抽出にも当てはまります。結果／抽出値のコンテクスト判別には、オントロジーが非常に有効です。」

「TDM により、信頼性と正確性の高い数値／定量的情報（投与量など）が見つかり、メタデータとしても抽出できます。これはその他のパラメータの抽出にも当てはまります。結果／抽出値のコンテクスト判別には、オントロジーが非常に有効です。」
ある大手製薬会社の情報科学者

TDMプロジェクトの例

TDM プロジェクトを設計する際の最大の課題の 1 つに、どのリソースやデータエレメントを取り込むかを考え、できる限り拡張させなければならない点が挙げられます。以下に、TDM テクノロジーの多岐にわたる使用事例をご紹介します。

- 本物の編集サポートと査読を提供すると称しながら、約束したサービスを提供せず、料金だけを著者や発表者に請求する、ハゲタカ出版社やハゲタカ学会を特定したい場合。学会発表者の組織数を図式化し、その毎年の数を信頼性の高い学会と比較したり、疑わしいジャーナルの著者や学会発表者の引用またはレファレンス指標を、他の著者や発表者と比較することで問題のある組織を検出することができます。
- 内部スタッフの情報探索強化のため、関心のある論文を読み、さらに学びたい研究者向け API の開発が可能です。API は、論文の DOI（デジタルオブジェクト識別子）を取り込み、他のスタッフが書いたそのトピックの論文に対し、図によるマッピング、関連するデータセットへのリンク、このトピックの研究助成金情報、著者が発表を行った学会へのリンクなどを読者に示すことができます。
- 大学図書館は、利用者ため、関連するデータセットや研究データをキュレートしたコレクションを作成することができます。Dryad などの既存の学術出版物の集合的データリポジトリのデータエレメントや、Springer Nature などの出版済みの書誌データベースのデータエレメントを組み合わせることで、インフォプロは研究者が特に関心を寄せる研究トピックについて利用可能なデータセットがないか、文献をモニタリングすることができます。関連するアプリケーションでは、データセットのレファレンスが他にないかモニタリングしたり、大学研究者によるデータセットへの全てのレファレンスを追跡することができます。

インフォプロは、もっとも権威のある、費用対効果の高い最良のオンラインリソースを所属組織にもたらす上で重要な役割を果たしており、組織内の TDM プロジェクトにも独自のスキルと専門知識をもたらすことでしょう。

Springer Nature は最近、TDM ツールやリソースについての情報のほか、TDM の使用法に関する Springer Nature ポリシーを掲載したポータルを作成しました。**springernature.com/jp/landing/text-and-data-mining** をご覧ください。

お問合せ
シュプリンガー・ネイチャー　インスティテューショナル・マーケティング

大学・政府機関のお客様
jp.market@ springernature.com

企業・病院のお客様
rd.japan@springernature.com