菌類学文献に隠された「未知の繋がり」を見つけ出す

中島 淳志

【背景・目的】

菌類学文献の全てに目を通すのは不可能である。単純に、あまりにも量が多過ぎるからである。筆者は１日３本の論文を読む習慣をおよそ７年続けているが、さながら怒涛のごとく流れ落ちる滝の水をコップ一杯で受け止めんとするような愚行で、仮にこれを一生続けたとしても、菌類のことを「知り尽くす」に至ることは決してありえないという確信がある。その上、一般論として、科学文献の量は顕著な増加傾向にある。菌類学分野の計量書誌学的研究は、筆者が渉猟した限りでは見出されなかったが、PubMed で「fungi」が中心的主題とされた文献を検索すると（"fungi[Majr]"）、平均ヒット数は 1990-1999 年で 4,177 件、2000-2009 年で 7,921 件、2010-2019 年で 11,646 件と、やはり右肩上がりの傾向が認められた。PubMed は主に医学文献のデータベースであり、菌類分類学の学術誌で収載されているのは「Mycologia」「MycoKeys」「Studies in Mycology」など一部であることから（「Mycoscience」は１件しかヒットしない！）、全容の把握は困難であるが、いずれにしても、生身の人間が日々追い切れる文章量ではなく、誰もが最新の集合知にアクセス可能な環境の実現には程遠いのが現状である。

したがって、膨大な知識を機械的処理により自動で整理し、ユーザーのニーズに沿った的確な情報源を提示するシステムが望まれる。全文検索はその一手段であるが、再現率（網羅性）が高い一方で適合率（精度）が低く、本当に必要な情報が埋もれてしまうおそれがある。検索者の意図を正確に汲み取るためには、Google がナレッジグラフの導入など様々な工夫により追求しているように、物事の意味を踏まえた検索（セマンティック検索）が必要であるが、その実現のためには、テキストデータから抽出した情報を適切な構造に整える必要がある。

主語（s）、述語（p）、目的語（o）の３点セットをセマンティックトリプル（あるいは単にトリプル）といい、セマンティック・ウェブの主要な構成要素である RDF（Resource Description Framework）ではこの表現手法が採用されている。この標準化された手法によってトリプル同士を繋ぎ、グラフネットワークを構築することにより、物事の関係性を機械にとって判読可能な情報に変換することができる。例えば、「何がきのこを食べるか」というクエリ（問い合わせ）に対して、述語に「食べる」、目的語に「きのこ」を指定して検索することにより、主語の一覧（ヒト、ネズミ、トビムシ、アライグマ…）を提示することが可能であるが、これは全文検索では実現困難なことである。生物多様性関連の文献やデータをトリプルの形式に変換する方法は多数検討されており（Cui et al., 2011、Stucky et al., 2014、他）、千葉菌類談話会通信 35 号で筆者が紹介した

Encyclopedia of Life の形質データベース「TraitBank」も内部ではこの構造を採用しているが（Parr et al., 2016）、菌類学に特化した研究は見出されず、菌類の構造化データが他の生物と比べても極めて乏しいという現状を鑑みても（中島、2019）、菌類学文献を対象としたトリプル抽出およびその自動化・効率化を試みる価値はあると考える。

また、知識をトリプルの形式で表すことには別の効用もある。1980 年代に D・R・スワンソンはLiterature Based Discovery（LBD）とよばれる新しい仮説生成手法を開発した（定訳はないが「文献に基づく発見」の意）（Swanson, 1986）。これは、互いに素、すなわち引用-被引用等の繋がりがない２本の論文から未知の関係性を特定するという試みであった。換言すれば、公表された「既知」と「既知」の組み合わせから「未知」を見つけ出そうとしたのである。スワンソンは「レイノー病が血液の粘度上昇などを原因とする」という知識と、「魚油が血液の粘度を下げる」という知識を三段論法の要領で関連付け、「魚油がレイノー病の治療に有効である」という仮説を見出し、それはのちに臨床的に実証された。同様の関係は偏頭痛や心房細動でも実証された（余談だが、スワンソン自身もレイノー病、偏頭痛、心房細動に苦しんでいた）（Smalheiser, 2017）。本手法は医学分野を中心に検討および改良が進められ、スワンソンが論文のタイトルから目視で関係性を見出したのに対し、近年では情報技術の進歩に伴い、検索対象は全文、抽出手段は機械学習、といった飛躍的なスケールアップも達成されている。Google Scholar で調査したところ、スワンソン論文の被引用論文には菌類学の論文は含まれておらず、この手法もまた菌類学分野においては未開拓と考えられた。

【方法】

2019 年 12 月から翌１月にかけて、PMC から Entrez 経由で「MycoKeys」など菌類分類学専門 27 誌の 1,932 論文の本文データを取得した。また、それに筆者所有の書籍および論文由来のデータを加え、文分割に支障のあるピリオドや括弧の除去などの前処理を加えたのち、nltk のトーカナイザを用いて文単位に分割した。続いて各々の文を AllenNLP の OpenIE モデルにかけて、動詞と目的語の組み合わせ毎に１つずつのトリプル候補を抽出した。各トリプル候補には出典情報（PMCID、ISBN、または DOI）を紐づけた。

主語と目的語の両方に名詞句（nltk の RegexpParser における正規表現：（{<JJ>*<NN|NNS|NNP|NNPS>+}）を含み、かつ主語または目的語に「①菌類の主要グループを指す一般的名詞（fungi、mushroom、lichen、yeast など）」、「②MycoBank から得られた属以上の分類群名」、「③ピリオド+半角空白」のいずれかを含むという条件を満たしたトリプルのみを抽出した。③の条件は「*C. albicans*」のような属名の短縮記法への対応として設けた。以上の絞り込みにより、菌類または菌類の学名を含む１対以上のエンティティを何らかの動詞により関連付けている文の取得を企図した。述語の動詞は nltk の WordNet レマタイザにより基本形に合わせた。

【結果・考察】

本手法により、自動的に約 50 万件のトリプル候補を得た。最も多かった述語は「have」で 22,196 件あり、「use」8,983 件、「show」7,562 件がそれに続いた。定量的評価は困難だが、適切に抽出されているトリプルが多い印象であった。一方、主に学名や専門用語を動詞と誤認するなどトリプル抽出に失敗する例も多く、無意味なトリプルが大量に生成されるという問題も生じた。

こうしたノイズへの対処手段としては、動詞での絞り込みが有用とみられた。例えば因

果関係を調べたい場合、「cause」、「induce」、「affect」、「influence」などの作用を表す動詞を事前に選んでおき、それらで絞り込むことで精度の向上が可能であった。別の問題としては、トリプルが文脈から切り離されることにより、時に誤謬を生じる例があった（例えば「高温は菌類の生長を促進する」は概ね正しいが、当然度を超えると死滅する）。そのため、実際にトリプルを採用するにあたっては、必ず引用元を参照し、それが正しいかどうかをその都度吟味する必要があると考える。他には、論理的には正しいが特定の状況・条件にしか適用できない例（「本研究では標本が入手できなかった」「３種を新種として記載した」など）や、常識的に明らかで情報としての価値が低い例などもあったが、それらをいかに抑制するかは今後の課題である。

本稿では、トリプルデータベースの用例として、以下の３点を紹介する。

① 主語・述語・目的語による「繋がり」検索と可視化

既に述べた通り、主語、述語、目的語を任意に指定することで、単なる全文検索よりも物事の意味を踏まえた検索が可能である。例えば、「胞子形成を促進するにはどうすればよいか？」という質問に対し、目的語「sporulation」を検索することにより、「pieces of carnation leaf / be add / to induce sporulation（カーネーションの葉/追加される/胞子形成誘導のため）」や「atmospheric carbon dioxide concentrations / amplify / alternaria alternata sporulation（大気中の二酸化炭素濃度/増幅する/*Alternaria alternata* の胞子形成）」、「eutrophic streams / significantly increase / fungal biomass and sporulation of aquatic hyphomycetes（富栄養の水流/顕著に増加させる/水生不完全菌のバイオマスと胞子形成）」といった潜在的に有用なトリプルを提示することが可能であった。また、Web アプリケーション（Dash Cytoscape）上で指定した単語の関連語をネットワークグラフの形式で表示させることで、対話的に洞察を得ることも可能とした（図１）。

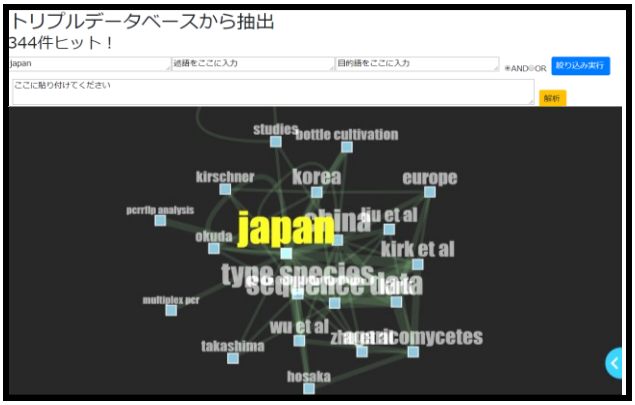

図１ Dash Cytoscape アプリの画面

② 任意のテキストにおける「繋がり」の可視化

作成したデータベースを利用する方法ではないが、同様の処理を任意のテキストに適用することにより、当該テキスト内での重要語の繋がりをネットワークグラフとして可視化することを可能とした。具体的には、Dash Cytoscape アプリに任意の論文の全文を貼り付けて「実行」ボタンを押すことで、ほぼ一瞬で図２のような結果を得ることができる。例えば Fukiharu et al.（2015）の全文から、当該論文で記載された新種の学名や、その種に深く関係する「エゾシカ」「糞」「日本」といった単語の繋がりを得ることができた。また、これを応用し、Entrez から日付指定で取得した新着論文を全てネットワークグラフ化して並べ、一目で通覧できるようにするとともに、その中から興味のある論文を迅速に選択することができるツールも検討中である。

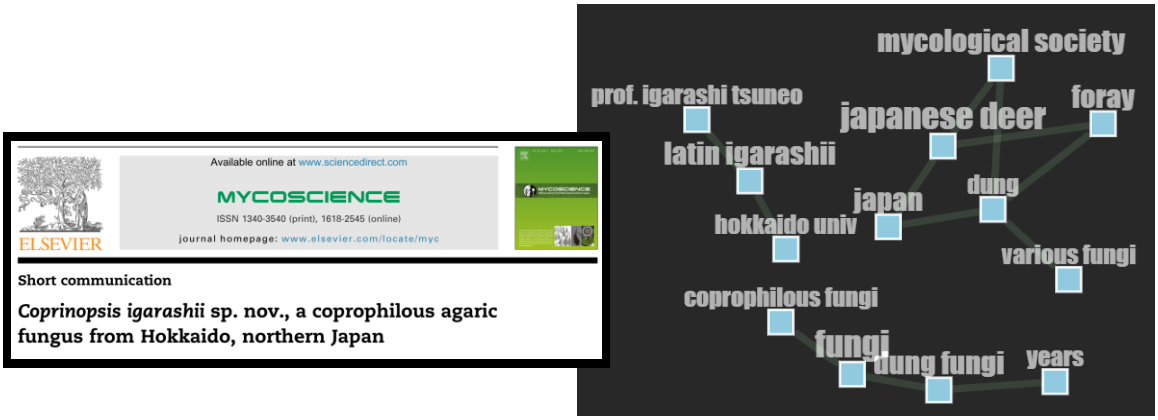

図２ 論文からのグラフネットワーク生成の例

③ LBD（スワンソンリンク抽出）

最後に、Dash Cytoscape アプリを用いた LBD の試行を可能とした。一見関連しそうもないと思われる任意の２つの語を選択し、それぞれを中心とするネットワークグラフにもし介在するノードがあれば、それを検討することで隠れた関連性を見出すことができる可能性がある。LBD には主に「オープン・ディスカバリー」と「クローズド・ディスカバリー」の２手法があるが、これは後者に対応する。図３は「競争（competition）」と「収斂進化（convergent evolution）」という、一見関連の見出せない２つの語の間に、「菌類（fungi）」と「オフィオストマトイド菌類（ophiostomatoid fungi）」の２ノードが介在することを示している。菌類学文献を基にしたデータベースなので前者は当然の結果であるが、後者については果たして真に関連しているかどうか、元データの精査も含めた検討の余地があるかもしれない。

また、下表には本手法で見出された、筆者が個人的に興味深いと考えた繋がりの一部を挙げる。こうした繋がりはあくまで仮説候補に過ぎず、それが真なのか偽なのか、未知なのか既知なのか、検証の価値が有るのか無いのか、といった判断は全てユーザーに委ねられる。現状では金属や魚群の探知機のようにある程度の確信を持った上で「ここ掘れ」と言えるものではなく、ほとんど途方もない宝探しに近い状態であるが、このような手段での仮説生成は「AI 時代の温故知新」の一つの姿と言えるのではないだろうか。

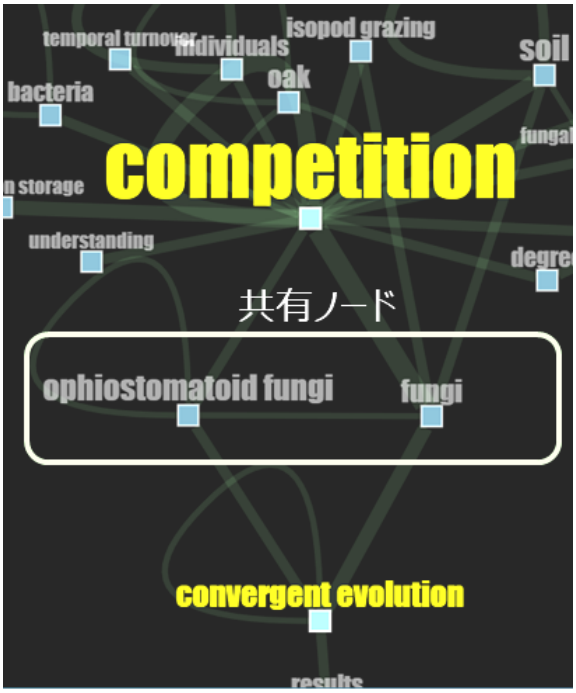

図３ LBD の結果の一例

表 本研究で生成したデータベースから得られた仮説候補の一部

免疫細胞→活性酸素→胞子の発芽
カドミウム→ストレス→マイコトキシン産生
植物防御応答→トリコデルマ→有性生殖
病原真菌→乾腐→建物の損傷
爪真菌症→B・カピタトゥス→発癌
絵画→A・プッルランス→過敏性肺臓炎
カリウム、塩素イオン流入→圧力上昇→胞子射出
ベータアミノ酪酸処理→T・ハルジアヌム→植物の抵抗性向上
トリコデルマ→エチレン産生→線虫誘引
青色光→ヒートショックタンパク質→分生子形成
シロイヌナズナ→揮発性有機化合物→分生子形成
植物のセスキテルペン類→アーバスキュラー菌根菌→植物の生長促進

【引用文献】

Cui, H., Jiang, K., & Sanyal, P. P. (2010). From text to RDF triple store: an application for biodiversity literature. Proceedings of the American Society for Information Science and Technology, 47(1), 1-2.

Fukiharu, T., Shimizu, K., Nakajima, A., Miyamoto, T., Raut, J. K., & Kinjo, N. (2015). *Coprinopsis igarashii* sp. nov., a coprophilous agaric fungus from Hokkaido, northern Japan. mycoscience, 56(4), 413-418.

中島淳志 (2019) 菌類の形質を把握せよ ～形質データの構造化・統制・集積～. 千葉菌類談話会通信, 35, 42-49.

Parr, C. S., Schulz, K. S., Hammock, J., Wilson, N., Leary, P., Rice, J., & Corrigan Jr, R. J. (2016). TraitBank: Practical semantics for organism attribute data. Semantic Web, 7(6), 577-588.

Smalheiser, N. R. (2017). Rediscovering don swanson: The past, present and future of literature-based discovery. Journal of Data and Information Science, 2(4), 43-64.

Stucky, B. J., Deck, J., Conlin, T., Ziemba, L., Cellinese, N., & Guralnick, R. (2014). The BiSciCol Triplifier: bringing biodiversity data to the Semantic Web. BMC bioinformatics, 15(1), 257.

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in biology and medicine, 30(1), 7-18.