２Ｆ０３

多様体学習による新たな研究動向手法の試行

○黒木優太郎（文部科学省科学技術・学術政策研究所）

1. 概要

ビブリオメトリクスを用いた研究分野のトレンドやホットトピックの抽出において、現在は、被引用数や共引用数、論文数を用いた手法が一般的である。しかしこれらの手法は、公開媒体（例えば論文データベース）の制約や、言語の壁が大きい。また、「数」を評価指標に用いることによって、そもそもの研究者コミュニティの規模に大きくバイアスがかかった結果を生み、「小さくても重要なトレンド」は埋もれてしまう傾向にある。

これらの問題に対しては、TF-IDF のように、数値補正する手法が一般的である。しかし本研究では、そもそも被引用数も論文数も使わず、言葉の文脈のみによってトレンドとホットトピックを抽出する手法を確立する。本手法は数を用いないため、研究規模は全く関係ない。媒体にも縛られないため、データ型の壁も存在しない。本報告では、試行的に英語論文を用いて「ゲノム」にまつわる学術トレンドを抽出した事例を紹介する。

2. 先行研究と現状

ビブリオメトリクスによる学術動向の調査では、被引用数や共引用などを用いた調査が主であるが、h-index を提唱した Hirsch, J. E.が既に自身で問題提起しているように、被引用数を用いた評価のみでは、研究規模のバイアスがかかる場合がある（Hirsch, J. E., & Buela-Casal, G., 2014）。そのため、例えるなら野球やサッカーのようなスポーツの重要性が取り上げられ、マイナースポーツのトレンドが埋もれがちであるという問題を本質的にはらむ。また、データ型に大きく依存するため、複数のデータを一つの手法で解析するにはコスト高になる。

また、自然言語処理を用いた研究動向把握の事例は多数存在するが、本研究と最も似た着想で、かつ有力な最新の報告としては、word2vec と TF-IDF を組み合わせたトレンド把握の実例が存在する（Hu, K., et al. "A Domain Keyword Analysis Approach Extending Term Frequency-Keyword Active Index with Google Word2Vec Model." Scientometrics, vol. 114, no. 3, 2018）。しかしながら、この場合にも、評価指標に「論文数」という頻度が混ざるため、根本的な問題を打破できていない。

その他にも word2vec を用いた解析は数多く行われているが、TF-IDF 等を用いて、論文数や被引用数などの何かしらの「数」の概念を使用してしまっているため、結局は研究規模に少なからず影響を受け、根本となるデータの制約を大きく受けることになる。ただし、上記手法には当然独自の強みもあり、精緻な意味地図を作成するなど、マッピングや単語の網羅的解析には適している。

一方で、特に科学技術予測における動向調査や、トレンド・ホットトピックの解析であれば、完全な網羅性は必要ない。そのため本研究は数の概念を捨て、形状による解析による動向把握に特化し、既存の問題解決を試みる。

また、「研究動向の把握」については、「科学技術予測」や「フォーサイト」に取り入れられており、日本では科学技術・学術政策研究所(NISTEP)が 1971 年以来半世紀以上にわたり実施している。また、NISTEP では、科学技術予測における「兆し」を見つけるための手法の開発も行っている（参考文献１）。昨今では、フォーサイトの重要性を各国が認知し、中国・韓国のような近隣国だけでなく、タイ、ロシアといった国でもフォーサイトが実施され、政策にも活用されている。

3. 結果概要と考察

本研究の手法の概要を図 1 に示す。まず、Scopus から「genome」でキーワード検索によって 2001～2018 年の論文約 320,000 報を得た。その後、それぞれの論文タイトルを対象に、多様体学習によってそれぞれの特徴語の抽出を行った。

さらに、これらをクラスタリングした結果、15 個のクラスターが得られた。 その後、これらのクラスターを 3 つのクラスに分類した。

クラス 1 :「科学」、「遺伝学」などの基本的な単語グループ。
クラス 2:ホットトピックまたはウィークシグナルグループ。
クラス 3 : 値が不安定なグループ。

これらのうち、クラス 2 について追加の文献調査をした結果、「microbiome」、「CRISPR / cas」、「Zika」、「colistin」などのいくつかの単語が実際に学術的に高く評価されていることが明らかになった。本研究によって、多様体学習とクラスタリングによって、研究動向が把握できる可能性が考えられた。

現状は英語論文による試行であるが、今後は、日本語論文や他分野の研究動向にも適応する。

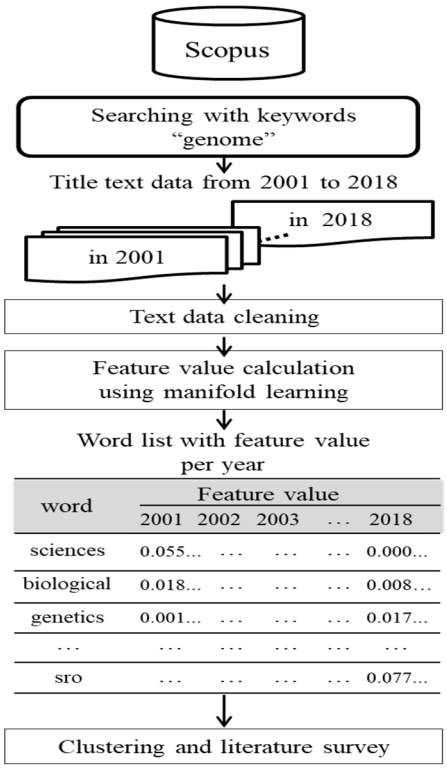

図 1 本研究の手法の概要

参考文献

[1] 科学技術予測センター，「兆しを捉えるための新手法～NISTEP のホライズン・スキャニング“KIDSASHI”～」，NISTEP Policy Study，No.16，文部科学省科学技術・学術政策研究所．DOI: http://doi.org/10.15108/ps016