様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 １２ 日現在

機関番号：８２６２４
研究種目：若手研究
研究期間：2020～2022
課題番号：２０Ｋ２０１４０

研究課題名（和文）数の概念を用いない、多様体学習に基づく研究動向解析手法の実証

研究課題名（英文）Research trend survey with simply utilizing manifold learning

研究代表者
黒木　優太郎（KUROGI, Yutaro）

文部科学省科学技術・学術政策研究所・科学技術予測・政策基盤調査研究センター・上席研究官

研究者番号：８０７４４３４１

交付決定額（研究期間全体）：（直接経費）　　1,900,000円

研究成果の概要（和文）：本研究では、Scopusデータベースから「ゲノム」に関する2001 年～2018 年までの320,000 件以上の論文を多様体学習で解析し、15 のクラスターに分類し、それをさらに3クラスに分類した。クラス①基本的な単語群のクラス。 クラス②注目キーワードクラス。「マイクロバイオーム」、「CRISPR/cas」など。 クラス③不安定なクラス。このうちいくつかの事例では、文献調査により、実際に学術的に評価の高い語句があったことを確かめた。
これにより多様体学習とクラスタリングのみを用いて、研究規模に依存せず注目キーワードを迅速に検出する新しい概念を提案した。

研究成果の学術的意義や社会的意義
学術的には、ビブリオメトリクスを用いたトレンド抽出においてトポロジーの概念を導入する。また社会的意義として、研究分野のビブリオメトリクスによる動向調査では、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は研究規模のバイアスがかかる。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。本研究によってこの問題を解消し、研究規模や媒体にとらわれず研究動向を分析可能とする。これにより、他者からの評価にとらわれない動向把握が、低コストで誰にでも可能となる。

研究成果の概要（英文）：In this study, more than 320,000 papers on "genome" from 2001 to 2018 from the Scopus database were analyzed by manifold learning, classified into 15 clusters, and further classified into 3 classes. Category 1: Basic word groups. Class 2: Featured Topics such as "Microbiome", "CRISPR/cas". Class 3: Unstable group. In some of these cases, a literature survey confirmed that there were actually words with high academic evaluation.
Thus, we proposed a new concept to quickly detect hot topics and weak signals without depending on research scale, using only manifold learning and clustering.

研究分野：科学技術予測

キーワード：研究動向分析　科学技術動向　多様体学習

１．研究開始当初の背景

研究開始当初の背景として、既存の研究分析ツールでは、医学や生物学といた巨大な分野の強みは良く見えるものの、経済学といった強みは把握しづらいという問題があった。このことは、国レベルの調査だけでなく、個々の大学といった現場レベルでも極めて重要な問題であり、最初の一石として、既存のツールや指標にとらわれない動向把握を検討するに至った。

そこで、昨今の自然言語処理の解析手法の内、特に単語のベクトル化、次元削減、多様体学習に着目し、「形」の可視化を行った結果、単語数が増えるにつれ、その形が線状に変化する事を見出した。よって、トポロジカルな形状変化に基づく動向把握を着想するに至った。

２．研究の目的

本研究では、これまでは多様体学習として t-SNE で行っていた形状変化の分析を改良し、より高精度に研究動向解析を行い、さらには日本語等の他言語でも同様の手法が実用可能であることの実証を目的とした。

具体的には、（１）個人の研究キーワードの同定、（２）研究キーワードの特定、（３）多言語対応の３つを目的とした。

３．研究の方法

（１）個人の研究キーワードの同定

事例として、ノーベル賞研究者のキーワード解析を行う。

（２）研究キーワードの特定

特定分野における研究キーワード分析を行う。特に、近年注目を集めている研究キーワードを抽出することを試みる。

（３）多言語対応

日本語で本手法が実装可能かどうかを確かめる。

＊本研究に関わる分野（科学技術予測、フォーサイト）における有力な国として当初はロシアがあったため、ロシア語を検討していたが、研究を進めている間に昨今の情勢変化があり、日本語のみとすることとし、英語手法の改善をあわせておこなった。

４．研究成果

（１）本研究を進める中で、t-SNE で行っていた多様体学習では１次元へ圧縮した際にグローバル情報が失われることがあり、指標化する際の精度に問題があることが判明した。そこで、新たにグローバル情報を保持する既存の多様体学習の適応を検討し、手法を改善した。

この改善によって、ノーベル賞受賞者のキーワード分析を行った結果、図１に示すように、医学や科学だけでなく、物理学や経済学といったノーベル賞受賞者であっても同様にグローバル情報を保持したままの次元削減結果が得られた。

実際、これらの受賞者のキーワード上位は、実際の受賞理由とも一部一致した。

ただし一致率はばらつきがあり Physiology or Medicine では約 90%な一方、Economic Sciences では約 70%であるため、さらなる改良を要する。

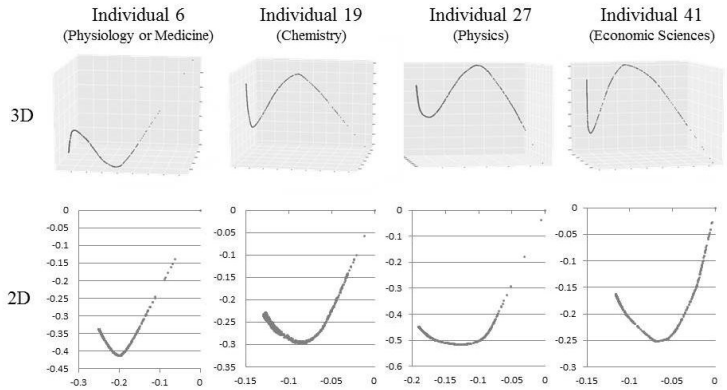

図１　ノーベル賞受賞者の論文の多様体学習による次元圧縮

（２）研究キーワードの特定

ある特定分野の研究キーワードを特定するため、論文データを分析した。論文データを分析するにあたっては試行的にゲノム科学を対象とした。さらにこれまでの手法を改善し、時系列の解析と可視化を可能とし、より高精度にキーワード抽出が可能となった。

具体的には、時系列での多様体学習と１次元への圧縮に加え、さらにそれらをクラスタリングすることで、いくつかのクラスタを得ることができた。クラスタ数はエルボー解析によって 15 と定めた。

クラスタ分析によって可視化することで、本手法における特徴量の変化にはくつかの累計が明らかとなった。

①基本的な単語群
常に特徴量の多いグループ。クラスタ１など。
「ゲノム」「遺伝子」等、当該分野において特徴的であることは間違いないが、常に上位であり、いわば当たり前のキーワード。
②注目キーワードグループ
ある時から接近したグループ。クラスタ４や５など。
何かしらの原因によって、この分野に接近したキーワード。注目度が上昇していると考えられる。
③不安定なグループ
常に低位のグループ。あまり重要でないキーワードか、エラーのグループ。

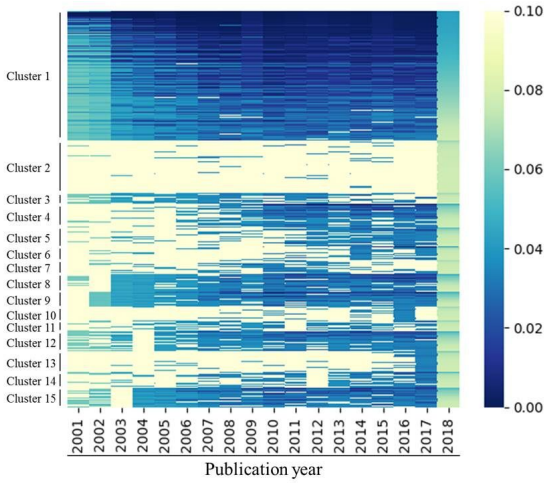

図２　ゲノム科学論文のキーワードクラスタ分析

グループ②のうち、特に急激に接近した（ゲノム科学分野における特徴が明らかとなった）キーワードについて、論文数を調べた結果、実際にその論文数も増加していること、それぞれの論文数の多さが一定ではないことがわかった。本手法によって、論文数の多い少ないに影響されず、その立ち上がりを検出したことが示唆された。

図３　グループ②の特徴的キーワードの論文数

（３）多言語対応
日本語でのデータ分析も行った。具体的なデータソースとして、まず（２）の結果を拡張する目的で、「ゲノム」をキーワードとして j-dream のデータを収集した。また、対象とする分野を広げ、さらに出版の形態に限らずに適応が可能かを検討するため、論文や書籍に限らずプレスリリースのデータを取得した。
これらの分析においては一定の成果が得られたものの、現状の手法ではデータ量が増えるごとに計算量が乗数的に増加するためクラスタ化に至らず、手法の改善を要する状態である。今後さらなる改良によって、日本語データにおいてもキーワード抽出を行う。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 黒木優太郎
２．発表標題 多様体学習による新 たな研究動向手法の試行
３．学会等名 研究・イノベーション学会
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関