「『特集名』特集号」　　解　　説

埋め込み法が拓くネットワーク科学の新展開

幸若 完壮*

1. はじめに

エクセルが登場するはるか昔、「クリミアの天使」と呼ばれたデータサイエンティストがいた。イギリスの看護師フロレンス=ナイチンゲールは、1854 年に勃発したクリミア戦争において、傷病者に関する膨大なデータから衛生状態と死亡の関連を突き止め、多くの命を救った。データから隠れた規則性を見つけ出し、次の行動に生かすことはデータ分析の大きな目的である。

従来、データは「ベクトルで記述されたモノの集まり」であった。例えば、傷病者に関するデータは、傷病者の特徴（身体的特徴や怪我の度合いなど）を記述したベクトルの集まりである。この「モノの集まり」を入力に想定して、これまで多くの統計分析・機械学習が開発されてきた。一方で、人間の言語やネットワークといった新たな形式のデータが登場してきている。

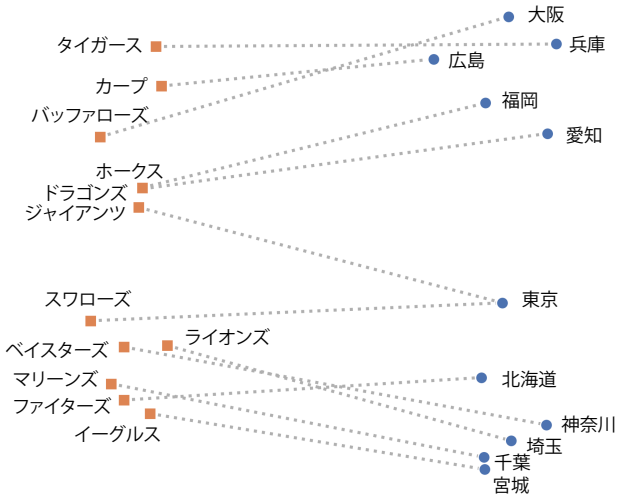
ネットワークは「モノの集まり」ではなく「モノのつながり」を表したデータである。友好関係、銀行の融資関係、Web など、ネットワークは身近に存在し、感染の拡大や倒産の連鎖など、社会に大きな影響を与える現象と深く関連している。これまで、ネットワーク分析のための様々な手法が開発されてきたが、データ形式の違いの問題で、一般的なデータ分析法の多くが十分に活用されてこなかった。

このデータ形式のギャップを解消する「埋め込み法」が登場し、ネットワーク分析に新たな流れを生んでいる。埋め込み法は、 頂点に座標を与え、位置で頂点の関係を表現する手法である。頂点の位置はベクトルで記述できるため、ベクトルを入力とする様々な分析法がネットワーク分析に利用できるようになる。

本稿では、ネットワークの埋め込み法と、埋め込み法が作るネットワーク科学の新しい流れを紹介する。具体的に、前半部では埋め込み法の仕組みを説明する。後半部では、埋め込み法を使って航空網と引用ネットワークを分析し、分析の流れや分析を助ける便利なツールを紹介する。

* Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, USA

Key Words: ネットワーク埋め込み法, 機械学習, 科学の科学



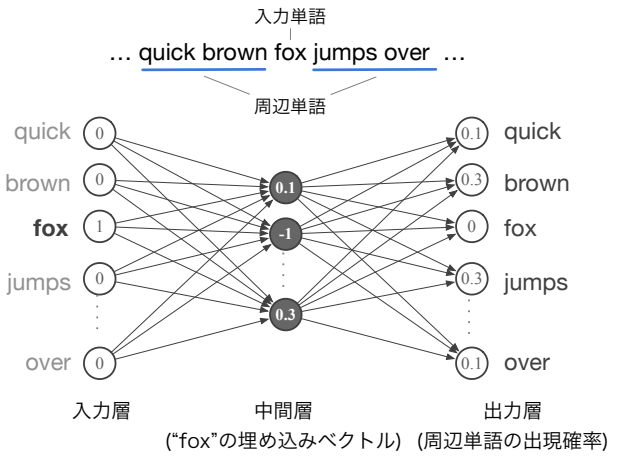
第 1 図　プロ野球球団と都道府県の分散表現。可視化のために埋め込み空間の次元 (300 次元) を主成分分析によって 2 次元に減らした。点線は各球団と本拠地の都道府県の対応を示している。この分散表現は Word2Vec に国立国語研究所の日本語コーパスを学習させて構築した [2]。

2. 埋め込み法

2.1 自然言語処理と埋め込み法

埋め込み法は、人間の言語を機械で処理しやすい形に変換する手法として発展してきた。率直に考えれば、1 つの単語に対して 1 つの数字を割り当てれば機械で処理できる形に変えることができる。例えば、「値段は高いが質が低い料理」という文を「値段, 高い, 質, 低い, 料理」と区切って各単語に 1 つの数字を割り当てて「1, 2, 3, 4, 5」と表現する。これは局所表現と呼ばれる単語の表現方法であるが、「高いの対義語は低いである」といった単語の関係を捉えられない。

この局所表現の欠点を解決したのが「分散表現」だ。分散表現は、各単語に座標を割り振り、単語同士の意味的な関係を位置関係で表現する。例えば、プロ野球球団と都道府県名を分散表現で記述しよう (図 1)。球団名と都市名を左右に分けて配置することで、それらのものが同種の単語であることが表現できる。さらに、単語をうまく配置すれば、簡単な線形代数で推論をすることができる。例えば、「北海道」と「ファイターズ」の単語から、「～ の球団」というベクトルを計算



第 2 図 Word2Vec は単語を入力としてその周辺の単語を予測するニューラルネットワークである。中間層が入力する単語の埋め込み空間上の座標である。

する。

$$v(\text{北海道})+v(\sim\text{の球団})=v(\text{ファイターズ}) \qquad (1)$$

ここで単語 w のベクトルを $v(w)$ とした。このとき、「福岡の球団は？」という問いの答えは $v(\text{福岡})+v(\sim\text{の球団})$、すなわち

$$v(\text{福岡})+v(\text{ファイターズ})-v(\text{北海道}) \qquad (2)$$

である。実際の日本語データから構築した分散表現では、この答えが $v(\text{ホークス})$ に大体一致する (図 1)。

2.2 Word2Vec

分散表現は膨大な単語の配置を決定する難しさや計算量が大きいといった実用上の課題のため実践が進まなかった。これらの課題を解決し、自然言語処理に大きな進展をもたらした手法が Word2Vec である [1]。図 1 の分散表現も Word2Vec から生成した。

Word2Vec は 3 層のニューラルネットワークである (図 2)。入力層は文章中の単語、出力層は周辺に現われる単語、そして中間層が入力単語の分散表現である[1]。

Word2Vec は情報圧縮装置と見ることができる。Word2Vec の入力は単語であるが、実際には単語を表す高次元のベクトルを入力する。このベクトルは 1 つの要素だけが "1"、他の全ての要素が "0" のベクトルで、"1" の場所で入力単語を表現する (図 2)。この高次元ベクトルは入力層から中間層に渡ってより小さな次元のベクトルに圧縮され、その後に中間層から出力層に渡って文脈を表すベクトルに変換される。この圧縮されたベクトルが入力単語の分散表現である。

Word2Vec では、似た文脈で使われる単語が近くなる傾向がある。異なる単語でも、使われる文脈が完全に同じであれば、出力層の周辺単語が一致し、出力層の信号元である中間層の分散表現も一致する。

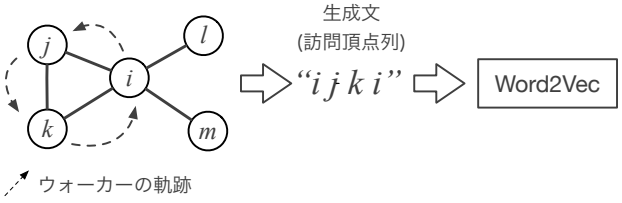
Word2Vec の興味深い活用例を 1 つ紹介しよう。研究 [4] では材料科学に関する約 330 万本の論文の要旨を Word2Vec に学習させ、様々な材料の分散表現を構築した。この「材料空間」は、様々な物性、例えば元素の周期表、物質の強磁性、熱電特性、結晶構造の類似性などを捉えている。また、材料の位置ベクトルを演算して、$v(\mathrm{Al_2O_3})-v(\mathrm{Al})+v(\mathrm{Si})=v(\mathrm{SiO_2})$ といった物質の反応や、$v(\text{二重六方最密構造})-v(\mathrm{La})+v(\mathrm{Cr})=v(\text{体心立方格子構造})$ といった結晶構造の関係を推論することができる。さらに、Word2Vec は過去の論文の要旨から当時は未発見の材料の特性を、2 から 4 割程度の精度で正しく予言したのである。

Word2Vec の登場によって自然言語処理は大きく進展した。Word2Vec はその後様々な改良が加えられ [5–8]、無償で利用できるパッケージにまとまっている [9]。

2.3 ネットワークの埋め込み法

自然言語処理の埋め込み法をネットワークに適用する手法を紹介する。自然言語処理の埋め込み法は文を入力とするため、当然ながらネットワークを入力に受け付けない。この文とネットワークという入力データの差を解消したのが DeepWalk である [10]。

DeepWalk では、頂点から構成される文をネットワークから生成する。例えば、ある頂点 i から文を生成するとしよう (図 3)。DeepWalk では、隣接する頂点 (j–m) のうち 1 つをランダムに選び、選んだ頂点を文に追加する。これを繰り返し行って十分に長い文を生成し、生成した文を Word2Vec に与えれば頂点の埋め込みが得られる。この「ある頂点から別の頂点を確率的に選んで移動する運動」はネットワークにおけるランダム・ウォークと呼ばれる。ランダム・ウォークには様々な種類があり、別種のランダム・ウォークを利用したネットワーク埋め込み法も提案されている [11,12]。



第 3 図 DeepWalk のしくみ。頂点を単語とする文をランダム・ウォークで生成し、Word2Vec に与える。

2.4 埋め込み空間の分析ツール

埋め込み空間は高次元であるため、直接理解することが難しい。ここでは埋め込み空間を分析して理解するためのツールを紹介する。

基本的な分析の方策は、頂点のメタ情報を活用して、埋め込み空間に解釈ができる軸を引くことである。例

[1] これは Skip-gram 法と呼ばれる。Word2Vec の別手法として、周辺語から中央の単語を予測させる CBOW(Continuous Bag-of-Words) がある [3]。

として、単語を埋め込んだ空間を考えよう。埋め込まれた単語の中には「楽しい」や「嬉しい」といった意味が肯定的な単語と、「つまらない」「悲しい」といった否定的な単語がいくつか含まれている。このとき、肯定的な単語のグループから否定的な単語のグループに向かって、単語の否定度合いを測る軸を作ることができる。これは SemAxis と呼ばれる手法である [13]。

SemAxis を設定するために、どのようなメタ情報が使えるのだろうか? これを探るために、線形の次元削減法である線形判別分析 (Linear Discriminant Analysis; LDA) が便利である。LDA は、ラベルが付いたデータ点を入力として、異なるラベルのデータがなるべく離れるように、低次元空間 (平面など) へデータ点を線形射影する手法である。メタ情報をラベルとし、LDA を用いて埋め込み空間を可視化することで、メタ情報と埋め込み位置の関連を調べることができる。

メタ情報がない場合でも、頂点の分布を可視化することで分析のヒントが得られることがある。埋め込みを可視化する代表的な手法として、主成分分析法、t-SNE[14]、U-map[15] などがある。いずれの次元削除法も頂点の位置を変えることに注意が必要である。例えば、可視化したときに距離が近い頂点であっても、元々の空間で近いとは限らない。そのため、可視化は分析の方針を立てる手段と捉えておくのが良い。

3. ネットワークの埋め込みの実践

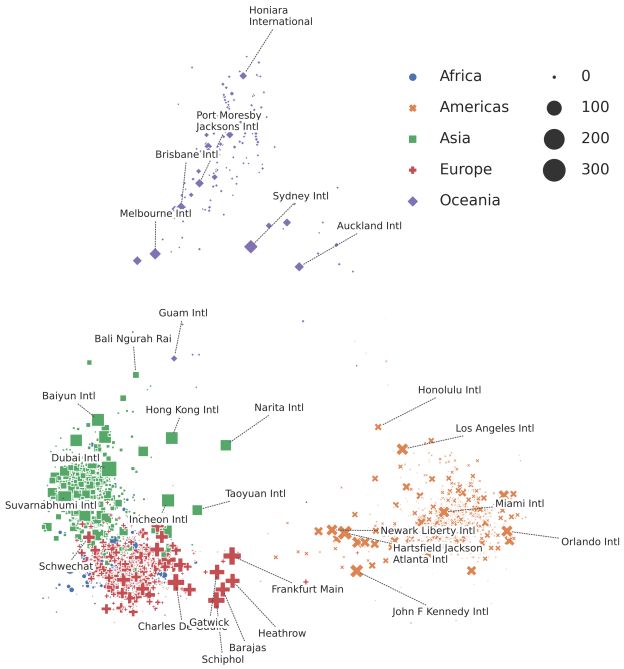
ネットワークはどう埋め込まれるのか？また、埋め込まれたネットワークをどう分析するのか？そして、ネットワークに埋め込み法を適用すると何が嬉しいのか？これらの問いを、2 つの実践例で答えていく。本稿で用いたコードは [16] で公開している。

3.1 空港ネットワークの埋め込み

空港ネットワークは空港を頂点、航空便の有無を枝で表したネットワークである。このネットワークを埋め込んだとき、空港の埋め込み位置と地理的な位置にどのような違いがあるか見てみよう。

空港ネットワークを構築するために、Openflight.org で公開されているデータ [17] を用いた。枝に方向と重みはない。このネットワークを、行列分解法に基づく DeepWalk[18] を用いて 128 次元の空間に埋め込んだ。

埋め込み空間では、空港の地域的な結びつきの他に、社会的・経済的な結びつきが表れている (図 4)。埋め込み空間には、アメリカ地域、オセアニア地域、そしてアジア・アフリカ・ヨーロッパ地域の空港からなる 3 つのクラスターがある。各クラスター内では地理的な近さだけでは説明できない空港の結びつきが表れている。例えば、オセアニア地域では、地理的にアジアに近いパプアニューギニアの空港 (Port Moresby Jacksons Intl) よりも、オーストラリアの空港 (Melbourne Intl, Syndney Intl) がアジア地域寄りに埋め込まれている。また、ヨーロッパ地域の空港は、地理的にはより近いアジアよりも、歴史的な理由で社会的・経済的な結びつきが強いアフリカ地域の空港と混ざっている。各地域クラスターの境界には、多数の地域を結ぶハブ空港 (Frankfurt Main, Heathrow, John F Kennedy Intl) や地域間を接続するゲートウェイ空港 (Guam Intl, Honolulu Intl, Narita Intl) があり、空港の位置関係から空港の役割を読み解くことができる。



第 4 図 空港ネットワークの埋め込み。可視化のため 128 次元の埋め込み空間を LDA を用いて 2 次元に射影した。各点は空港を表し、点の大きさはネットワーク上で隣接する空港の数を表す。LDA への入力ラベルとして、空港の地域 (アジア、アフリカ、アメリカ、オセアニア、ヨーロッパ) を用いた。

3.2 ライフサイエンス研究の埋め込み

新型コロナウイルス (COVID-19) の流行による社会・経済的な活動の制限が長期間続いており、治療法やワクチンの開発が精力的に進められている。

治療法やワクチン開発には、様々なプロセスを段階的に経る必要がある。例えば、ワクチンを開発するためには、ウイルスの性質を調べ上げ、ワクチンの候補を作る必要がある。次に、動物実験や臨床試験を行ってワクチンの安全性を確認しなければならない。さらに、開発したワクチンの大量生産や、市民に届けるための社会的な仕組みの構築などが必要である。この基礎研究から臨床応用までの段階的なプロセスは “Bench-to-Bedside”、つまり実験台 (bench) から患者の枕元 (bedside) までのプロセスと呼ばれている。

Bench-to-Bedside のプロセスは優れた基礎研究の成果を実用化するために重要である。では、基礎研究が

臨床応用にどのように橋渡しされているのだろうか？橋渡しの工程の中で、どの研究領域がどの程度重なっているのか？断絶はないか？これらの問いを、埋め込み法で答えてみよう。

優れた基礎研究は、臨床研究の土台として引用される [19]。この考えのもと、雑誌の引用ネットワークを埋め込み、埋め込んだ空間に基礎研究から臨床応用へのスペクトルがあるか確認する。雑誌の引用ネットワークを構築するために、ライフサイエンスの文献情報を収録した MEDLINE を用いる [20]。MEDLINE には 5,274 の雑誌が登録されており、専門家によって雑誌が分野別に分類されている。これらの雑誌の引用ネットワークを構築し、このネットワークを 128 次元の空間に DeepWalk を用いて埋め込んだ (図 5A)。

全体的に、同じ分野の雑誌が大体まとまって配置され、分野の位置から基礎研究と臨床研究へのスペクトルを見ることができる。具体的に、基礎研究である微生物学 (Microbiology) から始まり、薬学 (Pharmacology, Medicine) や腫瘍学 (Neoplasms), そして公衆衛生 (Public Health) へと分野が連続して重なっている。臨床研究が多い小児科学 (Child) や看護学 (Nursing) は、公衆衛生と近いが重なりは比較的少ない。

基礎研究から臨床研究までのスペクトルをより定量的に求めよう。微生物学の雑誌を基礎研究グループ、小児科学と看護学の雑誌を臨床研究グループとし、2 グループが極力混ざらないような軸を LDA によって求めた (図 5B)。この軸を基礎臨床軸と呼ぶ。この軸上での雑誌の順序は、専門家が行った雑誌の基礎・臨床の分類と一致している [21]。軸の両端の分野 (Microbiology, Child, Nursing) を橋渡しする分野には、一般科学、薬学、腫瘍学、公衆衛生がある。

基礎研究の成果は臨床研究の土台になっているのだろうか？この問いに答えるため、雑誌を基礎臨床軸上の位置でランキングし、10 のグループに等分してグループ間の引用数を数えた (図 5C)。全体的に、各グループは軸上で隣接するグループの雑誌を多く引用している、また、研究の段階が臨床研究に進むにつれてグループ間の引用が少なくなっている。引用が基礎から臨床に進むにつれて減少しているのは、臨床研究の論文が比較的少ないことが要因である可能性がある。引用の数を見ると、臨床から基礎への引用が 3 倍多い (臨床から基礎は 365,898,827 引用、基礎から臨床は 113,195,427 引用)。したがって、臨床研究は基礎研究の成果を元に展開されていることがわかる。一方で、基礎研究は、臨床研究の成果を取り入れながらも、他の基礎研究の成果をより多く取り入れて展開していることがわかる。

4. おわりに

ネットワークをベクトルに変換する技術、ネットワーク埋め込み法を紹介した。埋め込みによって、一般的な機械学習法や統計解析法を使ったネットワークの分析が可能となる。また、埋め込み空間における頂点の位置関係から、頂点の役割やネットワークにおける立ち位置を読み解くことができる。

実践例として航空網と雑誌の引用ネットワークを分析した。分析に用いたコードは [16] から入手することができる。雑誌の引用関係の分析に関して、全分野の雑誌の引用関係を埋め込み法で分析した研究 [22] を参考に、本稿では生命科学の分野に焦点を絞って分析した。埋め込み法を使って研究活動を分析する試みとして、学術用語を埋め込んだ研究 [23,24] がある。

本稿では DeepWalk を紹介したが、他にも LINE[25]、node2vec[11]、PTE[26] などのネットワーク埋め込み法が広く使われている。また、特種なネットワーク、例えば、2 部グラフを埋め込む方法 [27] や多重ネットワークを埋め込む方法 [12] も提案されている。新たな手法が提案される一方で、ネットワーク科学の視点で埋め込み法を見直す研究もある [29,30]。

埋め込み法は、簡単なベクトル演算でネットワークを分析する手段を提供している。一度埋め込んでしまえば、ネットワーク科学の知識がなくても、配置されている距離や方向から、関係の強さや種類を調べることができる。埋め込み法はネットワーク分析のハードルを下げただけでなく、ネットワーク科学と様々な分野の知見をつなぐ機会を作っている。統計がナイチンゲールによって看護に展開され、それまでの常識を大きく変えたように、ネットワーク科学の新たな展開が今後期待される。

参 考 文 献

[1] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proc. Conf. on Empirical Methods in Natural Language Processing*, 1389–1399, Lake Tahoe, Nevada, 2013.

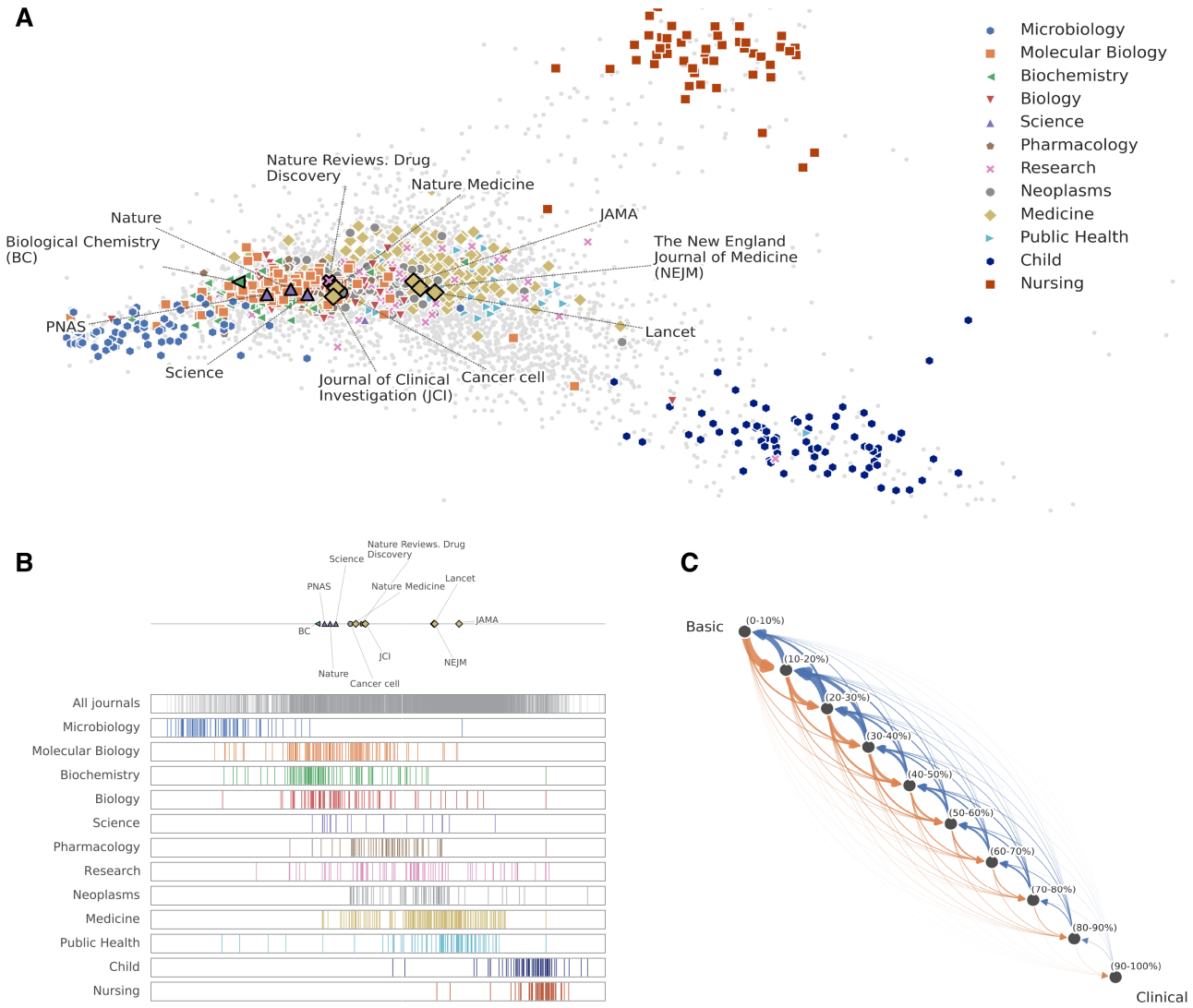
[2] 河村宗一郎, 久本空海, 真鍋陽俊, 高岡一馬, 内田佳孝, 岡照晃, 浅原正幸. chiVe 2.0: Sudachi と NWJC を用いた実用的な日本語単語ベクトルの実現へ向けて. 言語処理学会第 *26* 回年次大会, 6–16. 言語処理学会, 2020.

[3] T. Mikolov et al. Efficient estimation of word representations in vector space. *arXiv*, 1301.3781, 2013.

[4] V. Tshitoyan, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571, 95–98, 2019.

[5] O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. *Advances in Neural Information Processing Systems*, 3:2177–2185, 2014.

[6] J. Pennington, R. Socher, and C. Manning.



第 5 図　生命科学分野の雑誌の引用ネットワークの埋め込み。MEDLINE に登録されている 5,274 誌を 128 次元の空間に埋め込んだ。**(A)** 可視化のため LDA を用いて 2 次元面に射影した。LDA に与えるデータ点（雑誌）のラベルとして、National Library of Medicine による雑誌の分野別分類を用いた。**(B)** 基礎から臨床までのスペクトル。このスペクトルは、基礎研究である Microbiology と、臨床研究である Nursing と Child がなるべく離れるように、埋め込み空間上の全雑誌を直線に射影したものである。軸上の雑誌の位置の中央値で分野を降順に並べている。**(C)** 基礎と臨床研究の引用ネットワーク。 全雑誌を (B) の軸上の位置で順にならべ、全雑誌を 10 のグループに等分した。黒円は雑誌のグループを表す。グループ i から j への矢印は i から j への引用を表す。矢印の幅は引用数を表す。

GloVe: Global vectors for word representation. In *Proc. Conf. on Empirical Methods in Natural Language Processing*, volume 19, 1532–1543, Doha, Qatar, 2014.

[7] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov. FastText.zip: Compressing text classification models. *arXiv*, 1612.03651, 2016.

[8] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[9] R. Řehůřek and P. Sojka. gensim. https://radimrehurek.com/gensim/index.html. [Accessed 26th Sept. 2020].

[10] B. Perozzi, R. Al-Rfou, and S. Skiena. DeepWalk: Online learning of social representations. In *Proc. ACM SIGKDD Internal. Conf. on Knowledge Discovery and Data Mining*, 701–710, New York, New York, USA, 2014.

[11] A. Grover and J. Leskovec. Node2Vec: Scalable feature learning for networks. In *Proc. ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining*, 855–864, New York, NY, USA, 2016.

[12] Y. Dong, N. V. Chawla, and A. Swami. Metapath2vec: Scalable representation learning for heterogeneous networks. In *Proc. ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Min-*

ing, 135–144 , Halifax, NS, Canada, 2017.

[13] J. An, H. Kwak, and Y.-Y. Ahn. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Proc. Annual Meeting of the Association for Computational Linguistics*, 2450–2461, Melbourne, Australia, 2018.

[14] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *J. Machine Learning Research*, 9, 2579–2605, 2008.

[15] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*, 1802.03426, 2018.

[16] S. Kojaku. https://github.com/skojaku/graph-embedding-review-ja.

[17] Why anchorage is not (that) important: Binary ties and sample selection, 2011. https://toreopsahl.com/2011/08/12/why-anchorage-is-not-that-important-binary-ties-and-sample-selection/. [Accessed 26th Sept. 2020].

[18] J. Qiu, Y. Dong, H. Ma, J. Li, K. Wang, and J. Tang. Network embedding as matrix factorization: Unifying DeepWalk, LINE, PTE, and Node2vec. In *Proc. Conf. on Web Search and Data Mining*, 459–467, Marina Del Rey, CA, USA, 2018.

[19] G. M. Weber. Identifying translational science within the triangle of biomedicine. *J. Translational Medicine*, 11, 1–10, 2013.

[20] MEDLINE/PubMed Data. https://www.nlm.nih.gov/databases/download/pubmed˙medline.html [Accessed 22th Oct. 2020].

[21] F. Narin, G. Pinski, and H. H. Gee. Structure of the biomedical literature. *J. American Society for Information Science*, 27, 25–45, 1976.

[22] H. Peng, Q. Ke, C. Budak, D. M. Romero, Y.-Y. Ahn. Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. *arXiv*, 2001.08199, 2020.

[23] M. Chinazzi, B. Gonçalves, Q. Zhang, and A. Vespignani. Mapping the physics research space: A machine learning approach. *EPJ Data Science*, 8, 1–18, 2019.

[24] Q. Ke. Identifying translational science through embeddings of controlled vocabularies. *J. American Medical Informatics Association*, 26, 516–523, 2019.

[25] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. LINE: Large-scale information network embedding. In *Proc. Conf. on World Wide Web*, 1067–1077, Republic and Canton of Geneva, Switzerland, 2015.

[26] J. Tang, M. Qu, and Q. Mei. PTE: Predictive text embedding through large-scale heterogeneous text networks. In *Proc. the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1165–1174, New York, NY, USA, 2015.

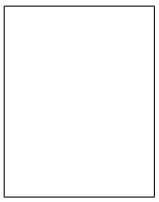
[27] M. Gao, L. Chen, X. He, and A. Zhou. BiNE: Bipartite network embedding. In *Internat. ACM SIGIR Conf. on Research & Development in Information Retrieval*, 715–724, New York, NY, USA, 2018.

[28] L. Meng and N. Masuda. Analysis of node2vec random walks on networks. *arXiv*, 2006.04904, 1–25, 2020.

[29] C. Seshadhri, A. Sharma, A. Stolman, and A. Goel. The impossibility of low-rank representations for triangle-rich complex. *Proc. of the National Academy of Sciences*, 117, 5631–5637, 2020.

[30] A. Tandon, A. Albeshri, V. Thayananthan, W. Alhalabi, F. Radicchi, and S. Fortunato. Community detection in networks using graph embeddings. *arXiv*, 2009.05265, 1–25, 2020.

幸 若 (こうじゃく)　完 壮 (さだもり) (非正会員)



2015 年 9 月北海道大学大学院情報科学研究家博士課程修了。2016 年 4 月英ブリストル大学博士研究員。2020 年 2 月米インディアナ大学博士研究員となり現在に至る。計算科学、ネットワーク科学の研究に従事。