Vol.2017-ICS-186 No.6
2017/3/4

ネットワーク表現学習によるネットワークの成長の可視化

浅谷　公威[1,a)]　大知　正直[1]　森　純一郎[1]　坂田　一郎[1]

概要：集団の挙動の理解や今後の予測には集団の発展を時系列に理解することが有用である。学術分野の引用関係などのネットワークデータから、集団が進化の過程を抽出し描画する手法の開発が進んでいる。既存手法では、各論文を集約したクラスタ間の離散的な時間における推移や関係性を描画しているため、個々の論文に関する情報を得ることはできない。本論文では、連続的な空間内に各論文を一つの点としてプロットし分野が徐々に広がっていく過程を 2 次元空間に描画し、領域の成長・分岐・融合の様子を表現しながら個々の論文の位置を明確にする手法を開発した。本手法では、まず、ネットワーク表現学習で得られた潜在空間での論文領域の成長方向を検出しその方向からのずれをカテゴリとして定量化し、次に、その上で近隣領域への連続的な進化のみを抽出する。これらのプロセスにより、複雑なネットワーク構造から領域の進化にそった関係性のみを抽出することを可能とした。本手法を用いて太陽電池や Graphene などの活発に研究されている領域のデータの可視化を行い、そのアウトプットが学術分野の理解に有効であることを検証した。

キーワード：ネットワーク, 表現学習

Visualization of evolution of network based on network representation learning

Kimitaka Asatani[1,a)]　Ochi Masanao[1]　Junichiro Mori[1]　Ichiro Sakata[1]

Abstract: Understanding the evolution of the group is useful for understanding group behavior and predicting future behavior. From the network data such as citation relations in the academic field, a lot of methods to extract and draw evolutionary processes of group are proposed. Using these existing methods, we cannot retrieve information of individual papers because nodes are placed in discrete time and discrete clusters.
In this thesis, we proposed the visualization method that each node is plotted as a single point in a continuous space for observing the process of gradually expanding the field is drawn in a two-dimensional space. In this method, firstly, the growth direction of the article area in the latent space obtained by the network expression learning. Next, the deviation from that direction is quantified as a category. Then, we draw the continuous evolution to the neighboring region Only. Through these processes, it was possible to extract only the relationships along the evolution of the region from the complex network structure. By using this method, we visualized data of actively studied areas such as solar cells and Graphene and verified that the output is effective for understanding academic field.

Keywords: Network evolution, Network representation learning

1. Introduction

科学、音楽、経済、社会において、個々の要素となる論文、楽曲、特許、会話はお互いの相互に関係し合うことで全体として流行・進化が起こり体系が構成される。個々の要素のつながりが明確に定義されたデータとして存在する場合、その体系の理解にネットワーク構造を用いたクラスタリング [4] が有用に機能している。近年では、ネットワークの成長過程を可視化し、領域の誕生・分岐・他の領域と

[1] 東京大学
1-7-1, Hongo, Bunkyo-ku, Tokyo
[a)] asatani@gmail.com

の融合を観察する手法 [5], [11] の開発が進んでいる。それにより、コミュニティの発展や論文の引用ネットワークの発展などを過去に振り返って理解することが可能である。

数多くの要素が複雑に絡まるネットワークの成長を理解するには、情報を集約し一部のみを抽出する必要がある。既存研究におけるネットワークの成長の可視化手法 [5], [11] では、横軸を年単位の時間として離散化し、縦軸もネットワーククラスタリングで得られたカテゴリへと離散化して情報量を適正な範囲に集約する。そのうえで、関係が希薄もしくは時系列の連続したクラスタ間の関係性のみを抽出して、クラスタの時系列の発展を 2 次元に描画する。しかしながら、時間とカテゴリの情報を離散化して描画した結果には以下の問題がある。まず、要素となる各論文の位置関係が明確に把握することができない。例えば、 2 つの論文の中間に位置している論文はどちらかのクラスタに属することになる。また、年のはじめの論文と年度末の論文も同時期ととらえられる。もう一つの問題点は、描画された各年のクラスタはその年までに出版されたすべての論文を含んでいることである。そのため、ある年に出版された論文だけにフォーカスを当てることはできないと同時に、分野の収束の事象を結果からすぐに把握できない。

本論文では、より直感的にネットワークの進化を理解することを目的とし、ネットワークの数万以上のすべてのノードを時系列の進化に合わせて離散化されていない 2 次元空間にマッピングする手法を提案する。そのことにより、領域の発生、消滅、融合、派生などの現象を示しつつ、ネットワークの発展の流れのなかでの各論文の位置を明確に示すことができる。提案手法はネットワークの個々の要素が追加されるたびに徐々に領域を広げていくという、Adjacent possible[6], [7] な変化を仮定したものである。Adjacent possible とは S. Kauffman が提唱した生物の進化は隣接領域の可能な領域のみに進化するという考えで、人工物や社会構造の進化を捉えることに応用されている。

本手法を、太陽電池やグラフェンといった様々な分野の論文データセットに適用した。そして、領域の発生、消滅、融合、派生などの現象を理解可能な形で 2 次元空間にマッピングすることで、学術領域が徐々に発展していく様子を描画し、有用な知見を抽出できることを確認した。

2. 手法

仮に論文の引用先が 1 つのみだとした場合、論文引用ネットワークは Tree 構造となるため、リンクの重なりなく 2 次元空間上に徐々に Adjacent possible な領域へ拡大するノードを配置可能である。しかし、現実の複雑な構造をもつネットワークが進化する様子を 2 次元空間にマッピングしても意味のある情報を抽出することは難しい。

本研究では、複雑なネットワークから Adjacent possible な領域で関係し合う関係のみを抽出することで、徐々に領域が拡大すると考えた。しかしながら、Adjacent possible な領域は空間上にノードを embed した後に計算可能となる。このような問題を解決するため、各ノードの位置（成長方向、カテゴリ情報）をある程度正確に計算して初期のインプットとして入力する。

そのうえで、算出した成長方向・カテゴリ方向を各軸とする 2 次元空間上で、Adjacent となる各ノードの空間的に近く引用関係のあるノードの距離がさらに近くなるよう、各ノードの位置を embed しなおす。このようにして、様々な距離の引用関係から近い引用のみを検出してノードの位置を集約していくことで、ノード群の成長・分岐・融合などの現象をハイライトする。本手法の概略は以下のようになる。

2.1 問題設定

ネットワークの近隣のノードのみのリンクの距離を最小化するように描画する。距離 D_a 以下の近隣ノード以外のエッジの距離は D_a と上限を定めた上で、各エッジの距離の和を目的関数とし、それを最小化するように手法を考案する。定式化すると、$argmin(\Sigma min(d(v_i, v_j), D_a))$ となるような各ノードの分散表現 v を学習する。以下の手法は必ずしもそれを直接的に最小化するものではないが、この目的関数を念頭においたものである。

2.2 提案手法

2.2.1 ネットワーク表現学習

ネットワーク表現学習とはノードの構造から各ノードの位置を表現ベクトルとして算出する手法である。引用ネットワークより、ネットワーク表現学習手法である LINE[14] 用いて、128 次元空間にノードを Embedding する。

LINE において 1 次と 2 次の 2 つの近接性が定義されている.1 次の近接性はノードのペア間のリンクの有無をもとに計算される. ノード i と j が接続されている確率は、表現ベクトル v_i と v_j より、式.1 で算出され、その確率が実際の接続関係に近くなるように各ノードの表現ベクトルを算出する。また、2 次の近接性は同じ接続先を共有するノードの近接度が高いという仮定に基づいて定義される。引用ネットワークでは、全く同じ論文群を引用している論文同士は 2 次の近接性が高くなり潜在空間ないの同じ位置に Embedd される。式.2 において， 各ノードはベクトル u および u' で表現され、一時の近接性と同様に実際の接続関係と式.2 で算出された接続確率の誘導が最大化されるように各ノードの表現ベクトルが算出される。

$$P_1(v_1, v_j) = \frac{1}{1 + exp(-\vec{u}_i^T \cdot \vec{u}_j)} \tag{1}$$

$$P_2(v_i|v_j) = \frac{exp(\vec{u}_i^T \cdot \vec{u}_j)}{\sum_{k=1}^{V} exp(\vec{u'}_k^T \cdot \vec{u}_i)} \tag{2}$$

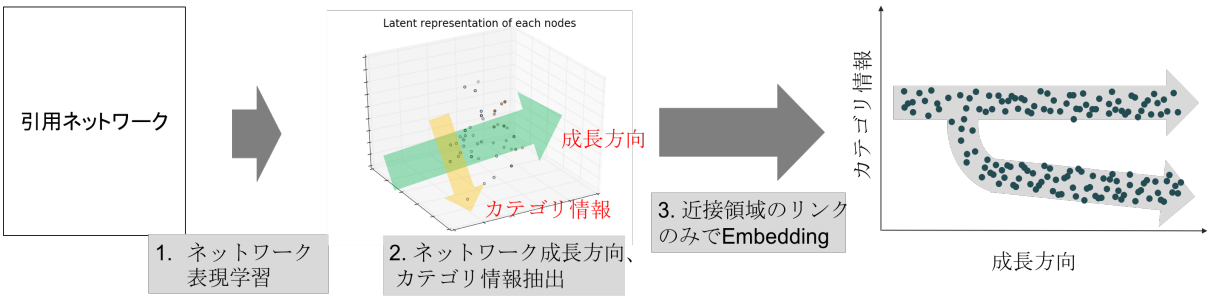

図 1　手法の概略

2.2.2　ネットワーク成長方向、カテゴリ情報抽出

上記の手法で得られた、ネットワーク表現学習により得られた分散表現空間上でネットワーク成長方向を抽象的な時間として算出する。そして、その方向性との差分ベクトルをもとに論文のカテゴリ情報とし、そのベクトルを TSNE[9] により 1 次元に次元削減する。

成長方向に関してはネットワーク表現学習で得られた潜在空間上で推定を行う。ノード生成時のみにエッジが作られる引用ネットワークの Latent Space 上での成長が方向性を持つことが分かっている [1]。この研究にもとづき、各論文の出版年を非説明変数、各論文の表現ベクトルを説明変数として重回帰分析とし、引用ネットワークの成長の方向性を算出する。そのモデルに各論文の表現ベクトルを入力して、各論文の引用ネットワーク上での擬似的な出版年（IPY: Intrinsic publish year）を算出する。

実際の出版年を用いず、成長方向から推定した擬似的な出版年を用いるのは、より正確に成長方向を描画するためである。例えば、30 年前の論文しか引用しておらず被引用もない昨年出版された論文が、同時期に出版された最新のトレンドを追った論文と同じ場所に配置されるのは適切ではない。

カテゴリ情報はその成長方向への差分として算出される。各論文から成長方向のベクトルへ降ろした垂線のベクトルをカテゴリ情報として算出する。こうして算出した各論文のカテゴリ情報の多次元ベクトルは、TSNE により 1 次元ベクトルとして次元削減される。

2.2.3　近隣進化のみを考慮した再 Embedding

上記の方法で得られた表現ベクトルは、横軸を成長方向 x、縦軸をカテゴリ情報 y として 2 次元空間内に配置される。その上で、各リンクのエッジ情報を考慮し、再び空間内で近隣領域のノードとのみ相互作用するようにノードの位置を変化させる。その様子を図 2 に模式化した。

再 Embedding の過程では、LINE の 2 次の近接性にならい、同じ接続先もしくは接続元ノードを共有するノード間で擬似的なリンクを作成し、そのリンクがつながれたノードどうしを空間的に近い位置に配置する。あるノード i の擬似的なリンクは、あるノード i の接続先を共有するノード群 B_i と、接続もとを共有するノード群 P_i から構成される。近隣領域のみからの成長を見るため、ノード i からのユークリッド距離が D_a より遠いノードは、B_i、P_i には含まれない。また、擬似的なリンクが接続するノード間の重みは、接続先を共有するノード間では (1/各ノードの出次数（引用数）の積)、接続元を共有するノード間では (1/各ノードの入次数（被引用数）の積)、として重み付けすることにより接続時数が極端に多いノードの影響力が極端に大きくならないように調整を行う。

これらのノード群と近い位置にノードのカテゴリ情報 y を以下の式に基づいて更新する。

$$y_i = \frac{\sum w_i \cdot y_j}{\sum w_j \, (j \in B_i)} + \frac{\sum w_i \cdot y_j}{\sum w_j \, (j \in P_i)} \tag{3}$$

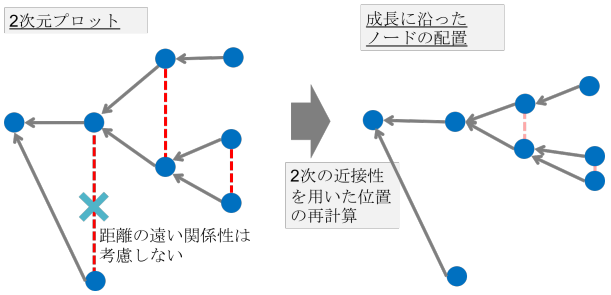

図 2　近隣進化のみを考慮した再 Embedding

3.　データ

様々な分野の論文情報をデータセットとして用いるため、公開されているなかで最も大きな論文データベースである Microsoft Academic Graph(MAG) を使用した。MAG は 2016 年 02 月 05 日時点での全学術領域の 1.2 億件以上の論文のタイトル・著者情報、カテゴリと、それらの間の 5.5 億件以上の引用関係を含んでいる。論文 1 件あたりの引用数は 5 件弱と若干疎であるものの、意味のある引用関係を抽出するのに十分である。

データベースはすべて Elastic Search DB に格納した上で、クエリ検索によって論文データセットを検索する。クエリは他の学術論文を対象とした分析 [12] で頻繁に使われているものから、ある程度引用関係がみつに取得できるものを選定した。クエリのリストは以下のようになる。

表 1 Datasets: citation networks.

Name	Query	#Nodes	#Edges
Solar cell	(solar cell or photovoltaics)	93923	1239979
Graphene	(graphene)	93923	1239979
Dopamine	(dopamine)	38825	374402

4. 結果

上記の手法を用いて、それぞれのデータセットを可視化した結果を以下にしめす。可視化結果をよりよく理解するため、Louvain 法 [3] を用いて引用ネットワークをクラスタリングし、各クラスタのタイトルを抽出して頻出語をリスト化した。

4.1 Graphene

Graphene 分野をネットワーククラスタリングした結果、表 2 のように複数の分野に別れる事がわかる。一番大きな分野は Yellow の色にあるクラスタである。このクラスタは、図 3 にみられるように、左の古い年代では大きな分野であったが、近年は収束していると考えられる。そこから、vapor film などのクラスタである水色のクラスタが派生している。近年では、青色のクラスタである ion やバッテリー関連の分野が発展していることが見て取れる。そこには、緑色のクラスタである light や photocatalytic のナノ分子に関する分野が融合していることがわかる。

以上の結果は、単純なクラスタごとの論文数の推移の可視化で観察できる部分もある。しかし、クラスタ同士が融合していく様子を明らかにした上で論文の各論文の位置を表示した手法には新規性があるといえる。クラスタの融合部にある論文、例えば緑色のクラスタの論文に関して詳細に個別の論文を観察することでどのように融合が進んだかを理解することができるかもしれない。

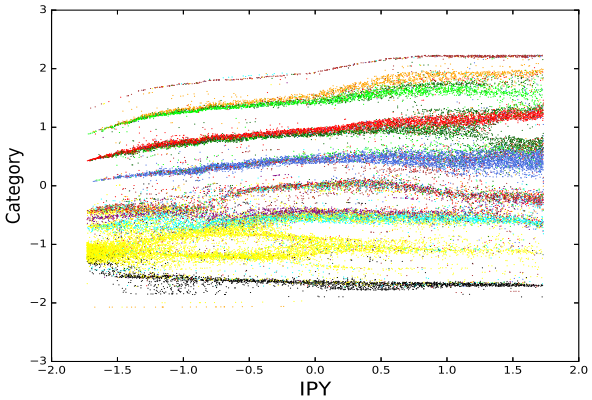

図 3 Graphene 領域の成長の可視化

表 2 Graphene 領域: クラスター一覧

Color	#Nodes	頻出語
yellow	11808	nanoribbons, properties, electronic, transport, bilayer, field, quantum
red	6256	oxide, properties, reduced, nanocomposites, composites, poly, polymer
royalblue	6119	oxide, lithium, high, performance, ion, batteries, synthesis, doped
darkgreen	3921	oxide, synthesis, photocatalytic, enhanced, nanoparticles, light
aqua	3914	growth, chemical, deposition, layer, vapor, films, carbon, synthesis
lime	3388	oxide, electrochemical, modified, electrode, reduced, nanoparticles, detection
black	2726	optical, laser, terahertz, fibe, tunable, surface, plasmon
orange	2307	oxide, detection, reduced, dna, functionalized
purple	1834	thermal, sheets, conductivity, molecular, mechanical, layer, nanoribbons
brown	1205	hydrogen, oxide, gas, sensing, adsorption,study

4.2 Dopamine

Dopamine 領域は複雑に領域が絡み合いながら成長していく様子が観察できる。近年では、赤の領域である parkinson 関連の領域が大きく進化しており、分野が実際の症例への応用に発展しつつあることが観察される。この分野には紫の領域である rat の脳の receptor に関する病気関連が融合してきており、また過去にこのクラスタには rat の recepter に関する領域が分離し進化していることが観察された。

細いラインで表示される小さな領域も存在している。これらの領域がどのように派生、分岐してきたかを分析していく必要がある。

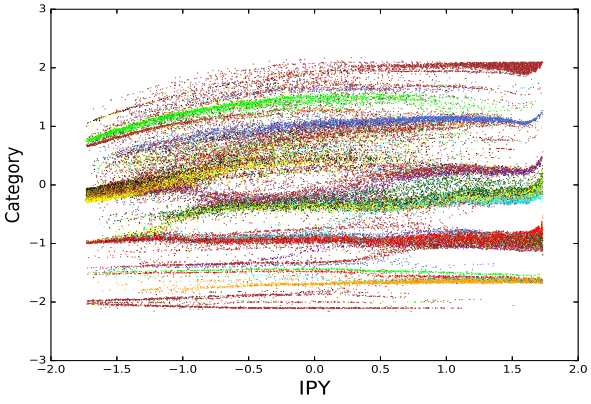

図 4 Dopamine 領域の成長の可視化

4.3 Solar cell

Solar cell 領域は、他領域に比べてかなり詳細な領域が独立に進化しているといえる。これは実際にこの領域では材

表 3　Dopamine 領域: クラスター一覧

Color	#Nodes	頻出語
yellow	8303	receptor, receptors, rat, d2
red	8143	disease, induced, parkinson, rat, neurons,dopaminergic
royalblue	7849	rats, nucleus, rat, induced, receptor,accumbens, neurons
darkgreen	4818	receptor, d2, striatal, receptors, schizophrenia, disease, pet
aqua	4435	renal, receptor, rat, receptors, cells,induced
lime	3531	cortex, prefrontal, receptor, rat, receptors, neurons, d1
black	3511	receptor, gene, association, d4, polymorphism, d2, transporter
orange	3391	transporter, cocaine, rat, release, uptake, brain, striatal
purple	2948	receptor, rat, induced, receptors, striatal,rats, disease
brown	2938	rat, release, brain, striatal, induced,rats, striatum

料別に研究が進んでいく様子がある。一方で分野の融合や分離という現象は、organic polymer(赤) や silicon film(青) といった、ネットワーククラスタリングで同一クラスタと判定された領域内で起こっている。これらの分岐がどのように起きて、分岐先がどのような分野になっているかはキーワード抽出などの手法で詳細に分析する必要がある。

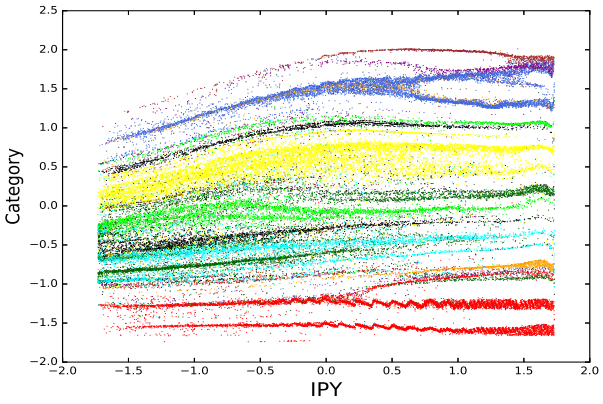

図 5　Solar Cell 領域の成長の可視化

5. 考察

様々なデータセットに本可視化手法を適用することで、分野の分岐、終焉、融合、発展といった結果を可視化することができた。既存手法と全く異なる方法で行うことで、より直感的な分野の理解が可能となったといえる。

また、本研究の結果は Adjacent Possible な領域に分野が発展するという考え方に基づいた分野の進化の理解が有用であることを示している。この Kauffman が提示した抽象的な概念の有効性を、大規模なデータから裏付けたのは一つの貢献であるといえる。

表 4　Solar Cell 領域: クラスター一覧

Color	#Nodes	頻出語
yellow	9722	power, systems, pv, grid, energy, connected, based
red	8965	organic, polymer, based, heterojunction, bulk, performance
royalblue	5139	silicon, film, si, light, amorphous
darkgreen	5042	silicon, si, efficiency, high
aqua	4475	dye, sensitized, tio, tio2, counter
lime	4421	film, cu, ga, se, cdte
black	3238	quantum, gaas, efficiency, junction, high
orange	2651	dye, sensitized, dyes, based, organic
purple	2247	dye, sensitized, perovskite, solid, state, based
brown	1214	quantum, sensitized, dot, cds

しかしながら、本手法の限界の一つはパラメータのセンシティビティである。結果において、比較的に人間が理解しやすいものを選択して描画した。どのような描画が適切であるかにかんして厳密な定式化が行うことが今後の課題と言える。もう一つは、分野の分岐・融合の一部を検出できていない点である。これは、2 次元空間内でノードの位置の再調整を行っていることに起因する。より高次元の空間内でのノードの位置を再調整し、適切に 2 次元空間に写像する方法を考案していく必要があると考えられる。

6. 先行研究

6.1 分野の発展の可視化

学術分野に限らず人工物の発展を理解することを目的とし、その対象とする要素に関するデータから発展する様子を時系列に可視化する方法が提案されている。各要素間のネットワークを分析する手法として提案されているのは [5], [11]、ネットワーククラスタリングにより得られたクラスタ間の推移を時系列に描画するものである。これらの手法の開発のモチベーションは、コミュニティを各時系列にどの順番に並べるかにおかれており、アクティブなコミュニティを上部に配置することや、時系列に異なるコミュニティ間のリンクのうち表示するリンクを選択する手法が提案されている。また、よりよいクラスタ間の時系列推移を観察する手法として、Rosva[11] らによる Bootstrap サンプリングによるある種のソフトクラスタリングによるクラスタ間の関係性の定量化などの発展しつつある分野である。

また、要素間のネットワーク構造を使わない方法として、LDA を拡張した離散的な時系列間でのトピックの推移を可視化する手法 (Dynamic Topic Model) も提案されている。

6.2 ネットワークの表現学習

本研究で使用するネットワークの表現学習とはネット

ワーク構造から要素（ノード）の分散表現を獲得する手法である．ネットワークの表現学習手法は 2014 年の DeepWalk に始まり数多くの手法 [2], [10], [14] が提案されており、既存の複雑ネットワークのクラスタリング手法よりもラベル推定や分類タスクを精度良く実施できることがわかっている。分散表現はテキストや画像を含んだヘテロジニアスなデータとの相性がよく、ヘテロジニアスなデータの分散表現化を行う手法の研究 [8], [13] されている。

ネットワークから得られた分散表現は TSNE[9] を用いることでローカルな構造を保ちつつ次元削減を行うことが可能である。様々なデータセットを対象にした実験が行われており、2 次元平面上に分離されたクラスタに各ノードが配置されるような可視化される。近年では、LargeVis という TSNE よりも高精度の可視化手法の研究も進んでおり、数百万のノードをクラスタに別れるように可視化することが可能である。

しかしながら、ネットワーク分散表現自体が空間的にどのようにマッピングしているかに関しては先行研究が少ない。我々は、ノード生成時のみにエッジが作られる引用ネットワークの Latent Space 上での成長が方向性を持つことを示した。この研究にもとづき、引用ネットワークの成長の方向性を算出し、引用ネットワーク上での各論文の擬似的な出版時期を定量化し、可視化に用いた。

7. 結論

本論文では、連続的な空間内に各論文を一つの点としてプロットし分野が徐々に広がっていく過程を 2 次元空間に描画し、領域の成長・分岐・融合の様子を表現しながら個々の論文の位置を明確にする手法を開発した。本手法では、まず、ネットワーク表現学習で得られた潜在空間での論文領域の成長方向を検出しその方向からのずれをカテゴリとして定量化し、次に、その上で近隣領域への連続的な進化のみを抽出する。これらのプロセスにより、複雑なネットワーク構造から領域の進化にそった関係性のみを抽出することを可能とした。本手法を用いて太陽電池や Graphene などの論文データセットの可視化を行い、そのアウトプットが学術分野の理解に有効であることを検証した。

謝辞

本研究は NEDO の委託事業「次世代人工知能・ロボット中核技術開発（次世代人工知能分野）」の一環として実施して得られた成果によるものである．

参考文献

[1] K. Asatani, O. Masanao, and J. Mori. Detecting research trend of academic field in latent space. In *First International Workshop on SCIentific DOCument Analysis (SCIDOCA 2016)*, 2016.

[2] S. Cao, W. Lu, and Q. Xu. Grarep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 891–900. ACM, 2015.

[3] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti. Generalized louvain method for community detection in large networks. In *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*, pages 88–93. IEEE, 2011.

[4] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[5] D. Greene, D. Doyle, and P. Cunningham. Tracking the evolution of communities in dynamic social networks. In *Advances in social networks analysis and mining (ASONAM), 2010 international conference on*, pages 176–183. IEEE, 2010.

[6] S. Kauffman, R. K. Logan, R. Este, R. Goebel, D. Hobill, and I. Shmulevich. Propagating organization: An enquiry. *Biology & Philosophy*, 23(1):27–45, 2008.

[7] S. A. Kauffman. Investigations. *Oxford University Press*, 2000.

[8] J. Leskovec. Beyond nodes and edges: multiresolution algorithms for network data. In *Proceedings of the 1st ACM SIGMOD Workshop on Network Data Analytics*, page 1. ACM, 2016.

[9] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[10] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

[11] M. Rosvall and C. T. Bergstrom. Mapping change in large networks. *PloS one*, 5(1):e8694, 2010.

[12] N. Shibata, Y. Kajikawa, Y. Takeda, I. Sakata, and K. Matsushima. Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. *Technological Forecasting and Social Change*, 78(2):274–282, 2011.

[13] J. Tang, M. Qu, and Q. Mei. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1165–1174. ACM, 2015.

[14] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077. International World Wide Web Conferences Steering Committee, 2015.