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

平成　３０　年　　６　月　１７　日現在

機関番号：　１２６０１
研究種目：　若手研究(B)
研究期間：　2016～2017
課題番号：　１６Ｋ１６１６７

研究課題名（和文）分野横断・融合的な新分野の成長予測と萌芽技術の特定

研究課題名（英文）Detecting emerging academic field by network algorithm.

研究代表者
浅谷　公威（ASATANI, KIMITAKA）

東京大学・大学院工学系研究科（工学部）・特任研究員

研究者番号：７０７７０３９５

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：本研究では、引用ネットワーク構造からトレンドを検出する方法を開発しました。科学技術政策や企業のR&Dのには、学問分野の動向を把握することが重要です。しかし、 従来の研究では学問分野のトレンドを数値化し、引用ネットワークから最先端の分野を検出することは困難でした。 本研究では、ネットワーク表現学習を用いた推論ネットワークの成長方向としての傾向を検出する新しいフレームワークを提案しました。 いくつかのデータセットにおいて、学問分野が潜在空間内で直線的に特定の方向に成長しトレンドが確認されました。そのトレンドは既存の方法と比較してより高い精度で将来の引用予測に使用できることを示しました。

研究成果の概要（英文）：In this research, we developed the method that detect academic trend from citation network structure. It is important to grasp trend of an academic field for the making science foresight and planning strategy of R&D. In previous studies, several network features and information retrieval methods have been proposed to elucidate the structure of citation networks and to detect important nodes. However, it is difficult to retrieve information related to trends in an academic field and to detect cutting-edge areas from the citation network. We propose a novel framework that detects the trend as the growth direction of a citation network using network representation learning. On several datasets, we confirm the existence of trends by observing that an academic field grows in a specific direction linearly in latent space. Moreover, we confirm that the detected direction can be used for future citation prediction with higher accuracy compared to existing method.

研究分野：　複雑ネットワーク

キーワード：　複雑ネットワーク　書誌情報

１．研究開始当初の背景

近年、学術文献の出版のペースや技術イノベーションの速度が加速しており、それらの最新の科学技術情報を用いて政策や戦略立案に関する意思決定を行うことが求められている。このような背景の元で、科学技術のロードマッピングやフォーサイト・ホライゾンスキャンニングのような意思決定支援のため、大規模な学術文献情報を分析するための技術やそのプラットフォーム開発に関する研究が、近年特に盛んに行われている。科学技術政策や企業の R&D のには、学問分野の動向を把握することが重要です。しかし、従来の研究では学問分野のトレンドを数値化し、引用ネットワークから最先端の分野を検出することは困難でした。

２．研究の目的

分野横断・融合的な新分野の成長予測と萌芽技術の特定を行う。具体的には分野のトレンドを数値化する手法を開発する。

３．研究の方法

本研究では、引用ネットワーク構造からトレンドを検出する方法を開発しました。本研究では、ネットワーク表現学習を用いた推論ネットワークの成長方向としての傾向を検出する新しいフレームワークを提案しました。本研究の貢献は、NRL の潜在空間上で引用ネットワークが一方向に成長する線形なモデル化が可能だと示したこと。それは言い換えると、引用ネットワークのトレンドの検知である。そのトレンドは、引用数予測などの将来予測に有用な指標となることを示します。

・トレンドの検知

近年の精度が高い Embedding 手法によって得られたノード の位置やノード間の距離は、これまでの手法以上に豊富な情報 を含んでいる可能性がある。もしそうであれば、ネットワーク上の任意の 2 つのノードの距離を定量的に算出することが可 能となる。算出された距離は、ネットワーク上の距離(相手の ノードに到達するまでの最短距離)よりも 2 つのノードの最短 経路上のノード以外のノードを含むという意味で豊富な情報を含んでいる。また、ネットワークをクラスタリングし所属クラ スタが同じか異なるかで任意の 2 つのノードの差異を測れるが、計測結果は 2 値となってしまい情報量が少ない。さらに、この距離の概念を用いると各ノードの性質を推定で きると考えらる。例えば、全てのノードへの平均的な距離が近 いノードはネットワークの中心に位置していると考えられる。論文の引用ネットワークで他の論文よりも距離が近い論文は古 くから分野の中心にある論文かそのような論文ばかりを引用し ている極めてクラッシックな論文である。逆に、他のノードか らの距離が離れた論文は Cutting Edge といわれるような既存 の分野から遠い論文であると考えられる。この類推が正しけれ ば、分野に新しい概念をもたらす”Cutting Edge”な論文が引用 されやすいという一般的な事実を組み合わせると、他のノード への平均かが長い論文は”Cutting Edge”であり引用数が伸ひびやすいと考えられる。これらの概念に組み合わせて、ネットワークの連続的な成長を考えるとネットワークが一方向に成長すると考えられる。

NRL を用いてネットワーク構造からマッピングされた潜在空間上の傾向（ネットワークの成長方向）を検出し、各ノードの傾向推移度を指標 IPY として定量化する。この研究で開発されたフレームワークを図 5 に示します。まず、いくつかのデータセットからデータを絞り込み、クエリを使用してサンプルデータセットを作成し、引用ネットワークを作成します。次に、NRL 法を用いて各紙の分散表現を取得する。 次に、研究動向（すなわち、学問分野の成長方向）を検出し、この方向に基づいて各論文の傾向推移の程度を推定する。最後に、IPY の特徴を定義し、各論文の将来の引用数を予測する。

直接引用と共引用ネットワークは、引用ネットワークの構造を解明するために重要である[30,31]。 LINE の代表的な学習方法（LINE 1st と LINE 2nd）は、直接引用ネットワークと共引用ネットワークに対応しているため、両方のネットワークの表現を作成するために、LINE [17]を使用します。

LINE 以外にも、過去数年間にいくつかの NRL 法が提案されている[20]。これらの方法は、行列分解法と、ノード間の近接関数を定義する方法とに大別される。いくつかの初期の方法[19,32]は、特定のラベル推定作業において他の方法よりも若干正確である。しかし、これらの方法はノード間の局所的な近接性を定義しないため、時間方向のネットワークの成長方向を検出するのには適切ではないと考えられる。近接関数は LINE と DeepWalk の両方で定義されている[18]。特に LINE 1st では、ノード間の近接関数はほぼ完全にエッジの存在に依存します。引用ネットワークにおけるノード追加の反復プロセスを考慮すると、新しいノードから既存のノードへの方向は、ノードが既存

のノードを引用するノードからの方向によって影響される。この繰り返しの間、引用ネットワークは特定の方向に成長すると推定される。

表現ベクトルの各次元において学問分野が特定の方向に成長していることを確認するために、LINE 第 1 行と第 2 行を用いて計算された 512 次元表現ベクトルの各次元の年平均を調べた。図 1 は各次元の表現ベクトルの年平均を示す。各次元の年平均値は、平均と標準偏差で正規化されます。図 1（a）は、特定の次元で直線的に成長する可能性のある各次元の平均値を示している。この傾向は、過去数年間に LINE 2nd（図 1（b））で観察された。用紙の位置が LINE 1st で直線的に変化する時間間隔は、各データセットの LINE 2nd よりも長くなります。 APS データセット（図 1（e））のふさわしくない線は、物理学の学問領域の多様性の結果と考えられている。全体として、NRL によって得られた各次元には成長方向に関する情報が含まれていることが確認された。具体的には、時間が進むにつれて用紙が各方向に直線的に特定の方向に移動される。この傾向は、データセットの最後の 5〜10 年の時間範囲ではっきりと観察されます。

・論文の空間へのマッピング
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。

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 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・他の領域との融合を観察する手法 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[5] 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。

４．研究成果
まずはじめに、。学術分野の引用関係ネットワークから分散表現を作成し、学術分野の成長方向を観測・予測するための技術を開発し た。 次に、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 Grap hene 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。 また、論文引用ネットワークを抽象化する基盤技術として、複数のレイヤーのネットワークから表現学習するための基盤技術を、人間 の移動データから場所の分散表現を作る技術をし実装することで実現した。
　論文の引用ネットワークの情報から技術のトレンドに関する知識の抽出をする手法として、ネットワーク表現学習により抽象化した分散表現と論文出版のタイミングから学術領域の成長の方向性を推定する手法を提案し、学術領域の成長方向と、成長のベクトルの先端（Cutting Edge）にある論文の引用可能性の高さを検証した。各論文のトレンドへの適合度を IPY(Intrinsic Published Year)として指標化し、その指標が PageRank などの従来手法よりも将来の引用数を高い精度で推定できることを確認した。各論文のトレンドを表す指標 IPY と分散表現をもとにして、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。異なる種類のエンティティ間の関係性を組み合わせて、エンティティの表現を学習する手法として、人間の移動データを対象に、場所間の移動ネットワークと場所間の地理的な関係性を組み合わせて場所の特性の学習を行った。既存の表現学習手法よりも高い精度でラベル推定を行えることを確認した。上記技術や、すでに学術産業技術俯瞰システムとして国家機関や企業へ提供している分析手法（論文群の俯瞰分析や萌芽論文検知）を広く多くのユーザへ提供するため、システムの改善にむけた基礎的な技術の構築を進めた。

また、上記の技術の開発と同時に多種データからのネットワーク表現学習の手法を提案し、Ubicomp 併設のワークショップで発表した。異なる種類のエンティティ間の関係性を組み合わせて、エンティティの表現を学習する手法を開発した。本手法では、人間の移動データを対象に、場所間の移動ネットワークと場所間の地理的な関係性を組み合わせて場所の特性の学習を行った。その際に、既存の表現学習を改良した手法を提案し、既存手法よりも高い精度でラベル推定を行えることが分かった。本手法は、論文の引用ネットワークとテキスト情報を対象とするなど、様々な領域に応用可能であると考えられる。

　上記技術や、すでに学術産業技術俯瞰システムとして国家機関や企業へ提供している分析手法（論文群の俯瞰分析や萌芽論文検知）を広く多くのユーザへ提供するため、システムの改善にむけた基礎的な技術の構築を行っている。まず初めに、論文データセットホルダである Elsevier 社と連携して Scopus データセットに対応したシステム設計を行った。また、Microsoft Academic Graph や Aminer.net が提供している ACM や DBLP といったドメインスペシフィックなデータから引用関係を抽出する技術を開発した。
　またシステムの実際の応用としてのアウトリーチの探索として、特定のドメインに関して詳細な分析を実施および、企業の分析担当者向けの講演を実施した。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 Da 以下の近隣ノード以外のエッジの距離は Da 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(Σmin(d(vi,vj),Da)) 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。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　1　件）

1. Kimitaka Asatani, Junichiro Mori, Masanao Ochi, Ichiro Sakata, Detecting trends in academic research from a citation network using network representation learning、 PLoS ONE (2018)

〔学会発表〕（計　3　件）

2. Kimitaka Asatani, Masanao Ochi, Junichiro Mori, Ichiro Sakata ,Predicting future citation from the temporal information of citation network, Second International Workshop on SCIentific DOCument Analysis associated with JSAI International Symposia on AI 2017,

3. Masanao Ochi, Yuko Nakashio, Yuta Yamashita, Ichiro Sakata, kimitaka asatani, Matthew Ruttley, Junichiro Mori Representation learning for geospatial areas using large-scale mobility data from smart cards, the 5th International Workshop on Pervasive Urban Applications in conjunction with ACM UbiComp 2016(PURBA2016)

4.Kimitaka Asatani, Masanao Ochi, Junichiro Mori, Detecting Research Trend of Academic Field in Latent Space, First International Workshop on SCIentific DOCument Analysis (SCIDOCA 2016)

〔図書〕（計　　件）

〔産業財産権〕

○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
浅谷　公威（ASATANI,　Kimitake）
東京大学・大学院工学系研究科・特任研究員
研究者番号：７０７７０３９５

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者

（　　）