Vol.2016-ICS-182 No.3
2016/3/2

Network Embeddingによる引用ネットワークの理解と予測

浅谷　公威[a)]　森　純一郎[b)]　坂田　一郎[c)]

概要：近年ではネットワーク構造から各ノードの表現ベクトルを推定する様々な Network Embedding 手法が提案されている．しかし，Network Embedding の応用先はラベル推定タスクやリンク予測タスクにとどまる．本研究では，論文の大規模な引用ネットワークをデータとして用い，Network Embedding が各ノードの性質の理解に有用であることを示す．具体的には，表現ベクトルを用いてノードのネットワーク内での位置や接続先ノードとの関係性を特徴量として定義し，その特徴量がノードの特徴を適切に捉えられることを論文の引用ネットワークを用いた分析により確認した．例えば，定義した特徴量である「周囲との平均的な距離」は，引用ネットワーク上で論文が”Cutting Edge”にあるか古い領域にあるかを表わしており，将来の引用数の増加が大きい論文はその特徴量が大きい傾向があることが分かった．また，将来の被引用数が大きい論文を推定するタスクにて，定義した特徴量および，既に提案されている表現ベクトルを使用したラベル推定手法が精度向上に寄与することが分かった．本論文で定義した特徴量は引用論文の推定のみならず，SNS 上の流行の伝播などの他の対象にも適用できると考えられる．

キーワード：Network Embedding, 引用数予測

Understanding citation-networks by network embedding

Asatani Kimitaka[a)]　Mori Junichiro[b)]　Sakata Ichiro[c)]

Abstract: Precision vector representation of nodes can be obtained by recent network embedding method. However, such vector representation is applied only for label prediction or link prediction. Our contribution is defining new features of networks using vector representation of each nodes and confirming that new features are useful for the understanding characters of each nodes and link prediction in paper citation networks. For example, from the feature ”average distance from node”, we can understand that the paper is in cutting edge or classical area. It is found that the feature of the growing paper, which acquires many citations in future, is relatively high compared with non-growing paper. And We confirmed that the feature improves accuracy of paper citation prediction task.

Keywords: Network Embedding, Predicting Citation

1. 序論

Network Embedding とはネットワーク構造をベクトル空間に写像することで各ノードの表現ベクトルを算出することである．近年の LINE[1]，DeepWalk[2] や Matrix Facrization[3] などの Network Embedding 手法を用いることで，ラベル推定やリンク推定などのタスクを精度よく行えることが分かっている．

Network Embedding の直感的の理解には，ネットワークの可視化に広く使われているバネモデル [4] を考えればよい．バネモデルはネットワーク構造を 2 次元空間に描画する一種の Embedding と考えられる．計算された各ノー

[1] 東京大学
7-3-1, Hongo, Bunkyo-ku, Tokyo, Japan
[a)] asatani@gmail.com
[b)] jmori@ipr-ctr.t.u-tokyo.ac.jp
[c)] isakata@ipr-ctr.t.u-tokyo.ac.jp

ドのベクトルの距離は任意のノード間の関係性への類推を与えてくれる．例えば，空間上の距離の遠いノードのペアは関係性が深くないように感じられる．しかし，バネモデルの写像は正確さを欠くため，上記の類推が成立するのは一部のペアに限られる．

このように，Network Embedding によって算出されるノード間の距離（近接性）は意味を持たないとされてきた．しかし，LINE や DeepWalk などはラベル推定タスクを高い精度でこなせることから，これらの手法で算出された表現ベクトルを用いたノード間の距離は意味を持つ可能性がある． しかし，表現ベクトルの距離がどのような意味を持つかに関しての考察は進んでいない．

Network Embedding を用いない既存の研究において，ノード間の距離は様々な方法で定量化されているが，それらはネットワーク上の一部の情報を用いているに過ぎない．例えば，ノード間の最短経路長はその経路以外の情報は参照されないし，Jaccard 係数 [5](接続先ノードの共有率) はノード間の周囲の情報のみを用いており，ノード同士の所属クラスタが同じかどうかという判別法 [6] も多くの情報を切り捨てている．Network Embedding は周囲の全てのノードとの関係性をもとに表現ベクトルを推定するものであり，それによって計算されるノード間の近接性は多くの情報を含んでいる．

そのような豊富な情報を含む Network Embedding 空間上での距離を用いると，あるノードのネットワーク上の特徴を理解できると考えられる．本研究では、ノードの「全ノードへの平均距離」や「接続先への平均距離」について考察した．「全ノードへの平均距離」が短いノードはネットワークの中心に位置しているだろうし，その逆の場合はネットワーク上に外れの位置にあると考えられる．また，「接続先への平均距離」，つまり，あるノードの隣接ノードが Embedding された空間で離れた位置にいるかどうかは，そのノードと周囲の関係性を推測する指標となる．例えば，隣接ノードの距離が近い場合はそのノードは特定のノードの密なクラスタ内にいると考えられるし，隣接ノードの距離が遠い場合は遠いクラスタのノードと多く接続していると考えられる．

本研究の目的は，Network Embedding を用いて「他のノードへの平均距離」および「接続先への平均距離」等の指標を定義し，それらがノードのネットワーク上での性質の理解や予測に有効であることを示すことである．

実験では，学術論文の引用ネットワークをデータセットとして用いた．学術論文をデータセットとして使用した理由は，ネットワークの成長を観察しやすいからである．引用ネットワークは新規に追加されたノードのみからリンクが貼られるネットワークである．従って，古いノード間にリンクが貼られることがないため，Embedding によって時系列に論文群が空間内でが成長する様子を確認できる．

定義した特徴量を用いて引用ネットワークを分析したところ以下の 2 つのことが分かった．初めに，成長論文（将来に一定期間内に引用数が増加する論文）は「他のノードへの平均距離」が長いことが分かった．「他のノードへの平均距離」はその論文が Cutting Edge にあるかどうかを表しており，そのような論文の性質を指標として定量化できたといえる．次に，定義した指標である「接続先への平均距離」を用いることで，成長論文は自身の分野以外の論文を多く引用している場合が多いことが分かった．

さらに，これらの特徴量が成長論文の推定タスクの精度向上に寄与することが分かった．その際には，既に提案されている，表現ベクトルを用いてロジスティック回帰よりラベル推定を行う手法も有効であった．成長論文の予測精度の向上は，科学技術の爆発的な増加からの知識の発見に寄与すると考えられる．近年では学術論文の出版数は爆発的に伸びており，コンピュータ・サイエンスや医学などの最先端の分野では出版から引用までの期間も短くなっている．そのような中で，計量文献学的手法により，将来的にコアとなるような論文を精度よく抽出することは，科学技術政策や企業戦略の立案に有用であると考えられる．

2. 先行研究

ここでは本研究が対象とする Network Embedding およびその応用および，目的タスクとなる成長学術論文の推定に関する先行研究について記載した上で，本研究の位置づけを明らかにする．

2.1 Network Embedding

これまで提案されてきた Network Embedding 手法として代表的なものは, Matrix Factorization[4] である．Matrix Factorization は n 個のノードのネットワーク行列 (n × n) を，$n \times m (m < n)$ の行列と $m \times n$ の行列の積に分解することで，各ノードの m 次元の表現ベクトルを推定する手法である．ラベル推定などで既存のクラスタリング方法に対して高い精度が出るには一部のタスクに限られているか，その応用の手法でも小さなデータセットに対してしか処理が出来なかった．

DeepWalk[2] は自然言語の Embedding 手法である Word2Vec[7] をネットワークに応用した手法である．Word2Vec とは文章から抽出された単語の並びから単語の性質をベクトル化する手法である．この手法では，表現ベクトルの内積が大きい単語は近くに共起しやすいという確率モデルを定義し，学習データとの尤度を最大化するようにニューラルネットを用いて表現ベクトルを推定していく．DeepWalk では，ネットワーク上をランダムウォークさせたときに得られるノードの時系列の並びを文章と見立てる．つまり，ノードが単語に対応し，ノードの並びが文章に対応している．Mikolov らは，そのノードの並びを

Word2Vec の学習器にかけることで，ノードの表現ベクトルを獲得することを提案した．

もう一つの代表的な Network Embedding 手法は LINE[1] である．LINE では 1 次と 2 次の 2 つの近接性を定義している．1 次の近接性は，直接繋がっているノードのペアの間で高いとされる直接的な繋がりを意味している．2 次の近接性とは，2 組のノード間で接続先のノードを共有している割合が大きいほど高いとされ，間接的な繋がりを意味している．LINE ではそれぞれの近接性定義に基づき，ASGD[8] により表現ベクトルを学習する．

LINE と DeepWalk を比較した場合，ラベル推定精度は後者のほうが若干上回る．DeepWalk における近接性の定義は SkipGram[9] モデル（前後の複数個の単語の共起頻度）に依拠しており、Line でいえば 2 次の近接性にあたる。しかし前者は、前後の複数の繋がりを参照する複雑なモデルであるが、LINE(2 次) は接続先しか参照しない単純なモデルである。さらに、LINE では 1 次の近接性も定義しているため、様々な観点からの分析が可能である。以上より，LINE のほうが結果の解釈の容易さや尺度の多さの面でリーズナブルな手法だと考えられる．従って本研究では LINE をベースとして解析を行う．

2.2 Network Embedding の応用

LINE や DeepWalk 等のここ数年の Network Embedding の応用タスクはラベル推定 [1], [2] である．ラベル推定は，YouTube や SNS や Wikipedia 等のネットワーク構造から，そのノードにつけられたラベルを推定することである．ラベル推定タスクの拡張として，多層ネットワーク（単語の共起・文書と単語・単語とラベル）から文章の分類を行う手法も考案されている．リンク推定タスクでは，Mao らが発案した Matrix Factorization で算出した表現ベクトルを用いたノードの抽象化手法 [3] をもとに，Song[10] らはこの着想より Matrix Factorization を用いて SNS 上の潜在的なリンクを既存手法より高い精度で予測することに成功している．

このように，Network Embedding はラベル推定やリンク推定のタスクに用いられているが，Embedding で得られる表現ベクトルを用いたノードの距離に関しての考察は進んでいない．本研究では，ノードから他のノードへの距離に着目したノードの指標化を行う．

2.3 引用予測

学術論文の書誌情報を用いて，成長論文（将来引用数が増加する論文）を特定する研究が数多く行われている．そのような分析は，引用ネットワークの特徴量を用いるものと，書誌情報を用いるものに大別される．前者の取組の代表的な例として，Barabasi らのグループによる引用数予測 [11] をはじめとした，様々な研究が挙げられる [12], [13]．後者は，雑誌や書誌の情報をそのままもしくは抽象化された情報を用いることで引用数予測に有効であることが示されている．しかし，書誌情報は分野によって異なるため汎用的な予測モデル作成には適さない．

成長論文の推定タスクにおいては，概ね出版後数年のデータを用いて出版 3 年後の引用数を予測している．本論文では，Embedding によって得られる特徴を明確にするために対象論文郡の被引用のない状態のデータを用いて分析を行い，3 年後の引用数を予測するタスクを実施する．

3. 新しい表現学習の定式化

LINE による Network Embedding 結果を用いて算出したノード間の距離や，ノード間の距離を用いて定義した特徴量についてここに記載する．

3.1 LINE

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同士がリンクの有無をもとに計算される．ノード i と j の表現ベクトル v_i と v_j を用い，式.1 に定義された確率でノード i と j が接続している確率が定義されている．式.1 より，v_i と v_j の内積が大きい場合に接続している確率が高くなる．式.1 で定義された $P_1(v_1, v_j)$ および, 実際の関係 $\widehat{P}(v_1, v_j) = \frac{w_{ij}}{W} (W = \sum w_{ij})$ (w_{ij} は i と j が接続している場合に 1，それ以外に 0) のとの KL 距離が小さくなるように，各ノードの表現ベクトル v を推測していく．

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

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いる．引用ネットワークで言えば，全く同じ論文の集合を引用している論文のペアは近接度が限りなく高くなる．一方で，接続しているノード同士のほうが近接度が必ずしも高いとは限らない．式.2 において，各ノードはベクトル $\vec{u}$ および $\vec{u'}$ で表現される．$P_2(v_i|v_j)$ はノード v_i から v_j へのリンクがある確率を表している．仮に，ノード v_j と同じ表現ベクトルを持っているノード v_k があった場合，ノード i に対しても $P_2(v_i|v_j) = P_2(v_i|v_k)$ となる．このように，2 次の近接性から求めた表現ベクトルが近いノードは，同様の接続先を共有している確率が高い．

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

2 次の近接性においても 1 次の近接性と同様に，定義された近接性と対応する実際の接続関係 $\widehat{P}(v_i|v_j) = \frac{w_{ij}}{d_i}$ (d_i はノード v_i の接続次数) と式.2 で定義された確率 KL 距離が近くなるように，各ノードの表現ベクトルを推測していく．

表 1: Data Sets: citation networks

	Query	Year	#Nodes	#Edges
Complex Network	TS=(social network* or random network* or complex network* or small-world or scale-free)	2014	158305	691661
Nano Carbon	TS=((carbon and (nano* OR micro*)) or fullerene or Buckminsterfullerene or Buckminster-fullerene or C60 or C-60 or graphene or (filament* and carbon))	2014	214168	2286648
Solar photo	TS=(Solar photo or photovoltaics)	2014	55957	149543

1 次および 2 次の近接性の双方において，SGD のアルゴリズムを改良し Lock 状態を回避した ASGD(asynchronous stochastic gradient algorithm) を用いて各ステップで KL 距離が短くなるように表現ベクトルを更新していく．

3.2 ノード間の距離

任意のノード間の距離は，1 次および 2 次の近接性の定義から算出したノードの表現ベクトルを用いて算出することが可能である．ノード間の距離はユークリッド距離で以下のように定義した．

$$d(v_i, v_j) = \sqrt{\sum_x (\vec{u_{ix}} - \vec{u}_{jx})^2} \tag{3}$$

3.3 ネットワーク特徴量の定義

1 次および 2 次の近接性の定義から算出したノードの表現ベクトルを用いて，各ノードのネットワーク上の位置を抽象化した特徴量である「全ノードへの平均距離」および「接続先への平均距離」を定義した．

3.3.1 特徴量 1: 全ノードへの平均距離

ノード v_i から全ノードへの平均距離 $DA(v_i)$ は以下の式.4 のように定義される．あるノード v_i から全ノードへの距離を平均したものを，全ノード間の距離の平均である $DA_{average}$ で正規化したものが $DA(v_i)$ となる．

$$DA(v_i) = \frac{\sum_{j=1}^{n} d(v_i, v_j)/n}{DA_{average}} \tag{4}$$

$$DA_{average} = \frac{\sum_{j=1\ j>k\ k=1}^{n} d(v_j, v_k)}{n * (n-1)/2} \tag{5}$$

全ノードへの平均距離 $DA(v_i)$ が短いノードはネットワークの中心に位置しており，その逆の場合はネットワーク上に外れの位置にあると考えられる．

3.3.2 特徴量 2: 接続先ノードへの平均距離

ノード v_i から接続ノードへの平均距離 $DC(v_i)$ は以下の式.6 のように定義される．あるノード v_i から接続先ノードへの距離を平均したものを，ノード v_i から全ノードへの平均距離 $DA(v_i)$ で正規化したものである．d_j はノード j の接続次数を表している．正規化に $DA(v_i)$ を用いた理由は，対象とするノード v_i が他のノードからはれた場所にいるか近い場所にいるかといった情報を差し引いて考えるためである．

$$DC(v_i) = \frac{\sum_{j=1}^{n} d(v_i, v_j) * w_{ij}/d_j}{DA(v_i)} \tag{6}$$

隣接ノードの距離が近い場合はそのノードは特定のノードの密なクラスタ内にいると考えられるし，隣接ノードの距離が遠い場合は遠いクラスタのノードと多く接続していると考えられる．

4. Embedding 結果および特徴量を用いた論文引用ネットワークの分析

本節では，論文の引用ネットワークから Embedding を用いて成長論文の性質の推定を行う．まずはじめに，LINE(1st) および LINE(2nd) の Embedding 手法別に, ネットワークの成長と成長論文がどのような領域に属しているかを 2 次元空間のプロットおよび全ノードへの平均距離 $DA(v_i)$ の指標を用いて確認する．その上で，全ノードへの平均距離 $DA(v_i)$ および接続ノードへの平均距離 $DC(v_i)$ という 2 つの特徴量を用いて成長論文の性質を確認する．

4.1 データ

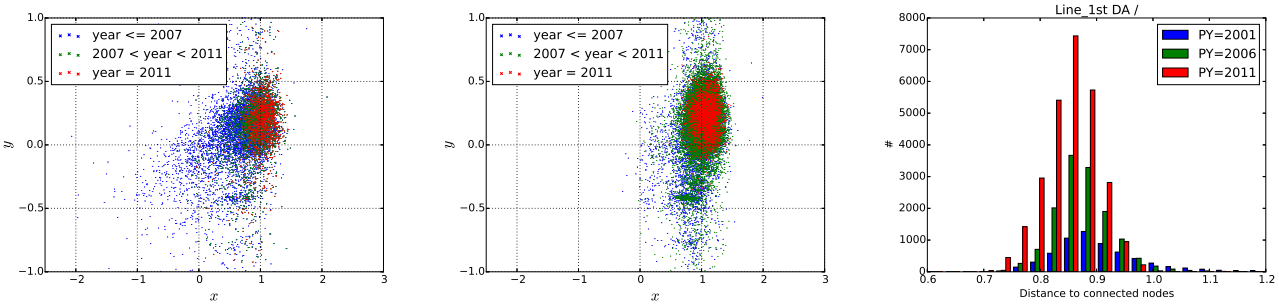
Thomson Reuters 社が提供してい Web of Science のデータベースをクエリで絞り込んだ特定の分野の論文の集合より引用ネットワークを作成して分析を行った．データの取得に使用したクエリと年度は表 1 のようになる．手法の汎用性を裏付けるため複数のデータセットを用い，データセットを作成するクエリは各分野の専門家に依頼した．

分析は各データセットについて行ったが，同じ傾向の結果が得られたため，一番大きなデータセットであるナノカーボンについての結果を以下に記載する．

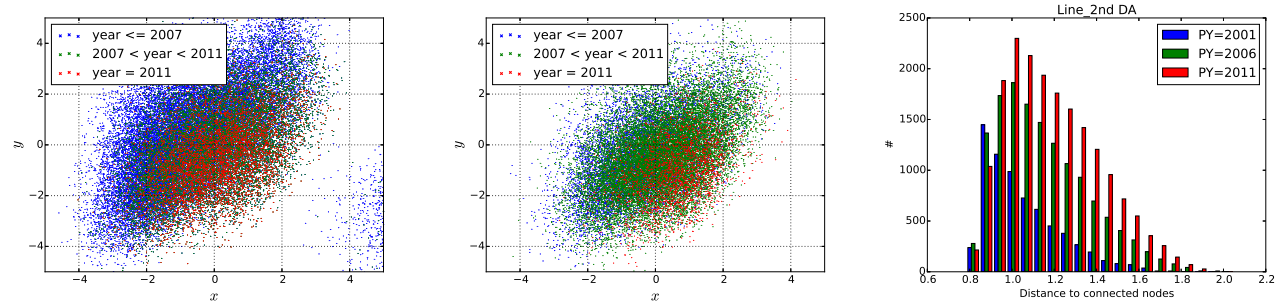
4.2 前処理: Embedding

データセットの年度 (2014 年) の 3 年前 (2011 年) までに出版された論文の引用ネットワークを使用した．

各データセットの論文の引用ネットワークより，LINE における 1 次および 2 次の近接性に定義に基づいた表現ベクトルを算出した．算出の際のパラメータは LINE の論文における実験設定と同様に，次元数 $N = 128$，イテレーションの繰返数 $S = 10000$，Negative サンプリングの割合



(a) LINE 1st - 年度別の論文領域　(b) LINE 1st - 2011 年の成長論文の領域　(c) LINE 1st - 出版年別頻度分布



(d) LINE 2nd - 年度別の論文領域　(e) LINE 2nd - 2011 年の成長論文の領域　(f) DA(LINE 2nd) 出版年別頻度分布

図 1: LINE を用いた論文引用ネットワークの可視化

$N = 5$ とした.

4.3 Embedding から見る引用ネットワークの成長と成長論文

論文の引用ネットワークの時系列の発展を観察するために，LINE(1st) および LINE(2nd) によって得られた Embedding 結果を PCA[14] 法により 2 次元空間へ写像した．2 次元に写像した結果は可視化にのみ使用し，特徴量の算出においては 128 次元のままのデータを使用する．

その結果を，2007 年以前，2008〜2010 年，2011 年の論文に分けて図 1a および図 1d にプロットした．その上で最新年度である 2011 年の論文の中で成長論文がどのような領域にあるかを調べた．2011 年の論文の集合を，2014 年度の被引用数を用いて，成長論文 (引用数増加が上位約 10%)，0-citation(今後 3 年で引用数の増加が 0)，その他の 3 つの集合に分割した．その 3 つの集合をそれぞれを異なる色で図 1b，1e にプロットした．図 1b，1e の縦軸および横軸は，図 1a , 1d と同じである．また，図 1c, 1f に DA(全ノードへの平均距離) の頻度分布をプロットした．

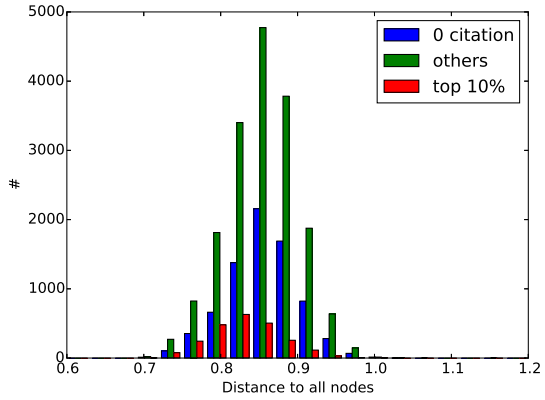
LINE(1st) による 1 次の近接性を用いた Embedding を行った場合，新しい論文が比較的新しい論文を引用することで年度が上がるたびに特定の方向に論文の領域が成長しているのとともに，古い論文への引用があることで新しい論文も空間の中心に引き寄せられる力が働くと考えられる．実際に図 1a において，2011 年度に出版された論文は，古い論文に比較して右側の領域に伸びつつあることが分かる．また，被引用数が多い論文は図 1b よりその年の出版年度の中でも中心によった論文であると考えられる．

LINE(2nd) による 2 次の近接性を用いた Embedding 場合，LINE(1st) と同様に年度が上がるたびに特定の方向に論文の領域が成長するメカニズムが働く．しかし，古い論文への引用があることで新しい論文も空間の中心に引き寄せられることはない．LINE(2nd) は引用関係がある論文同士が近いわけではなく，同じ論文を引用している論文が近い距離にあるからである．実際に図 1d において，2011 年度に出版された論文は，古い論文に比較して右側の領域に伸びているが中心に集まっている様子はない．図 1f は年度別に DA(全ノードへの平均距離) の頻度分布をプロットしたものである．この図より新しい論文が全体として中心から離れたりしていくことが確認された．また，被引用数が多い論文は図 1e よりその年の出版年度の中でもネットワークが成長している方向の最先端の論文 (Cutting Edge) であることが確認された.

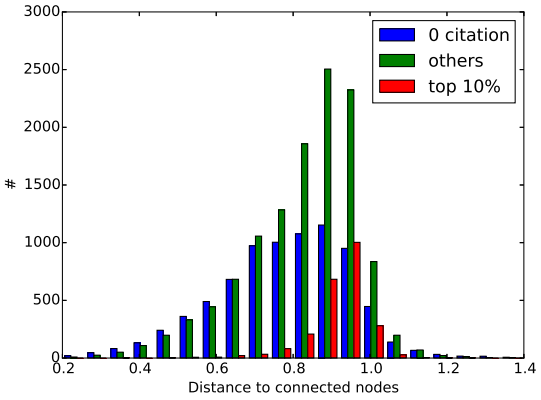
4.4 成長論文の性質

以上の分析の理解を深めるため，定義した指標である DA(全ノードへの平均距離) および DC(引用先論文への平均距離) を用いた分析を行った.

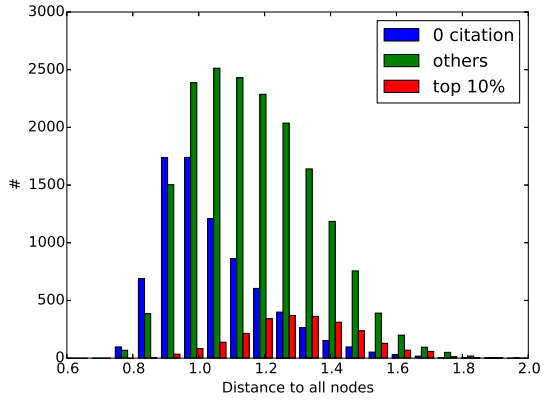
1 次の近接性 (LINE 1st) を用いて DA(全ノードへの平均距離) および DC(引用先論文への平均距離) のそれぞれの特徴量に対し，0-citation，その他，成長論文別に頻度分布を求めた．図 2a は，DA の頻度分布をプロットしたもの



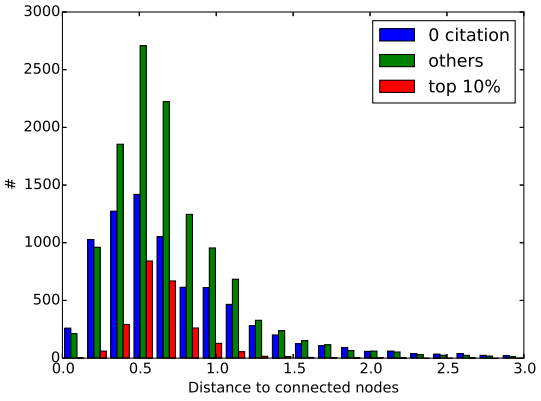
(a) DA(LINE 1st) 引用数別



(b) DC(LINE 1st) 引用数別



(c) DA(LINE 2nd) 引用数別



(d) DA(LINE 2nd) 引用数別

図 2: 0-citation，その他，成長論文別の定義した特徴量の頻度分布

であるが 3 つの論文集合において大きな差が見られなかった．その一方で，図 2b より DC(引用先論文への平均距離) は成長論文，その他，0-citation の順に大きいことが分かる．この結果より，1 次の近接性で Embedding した結果，成長論文は「遠くの距離にある論文を引用している」傾向があることが分かった．

2 次の近接性 (LINE 2nd) を用いて算出した特徴量を使って分析した結果は，1 次の近接性 (LINE 1st) を用いた結果と異なるものとなった．図 2c より，DA(全ノードへの平均距離) は成長論文，その他，0-citation の順に大きい傾向があることが読み取れる．これは，2 次の近接性で Embedding した結果，成長論文は「他の論文からの距離が遠い」傾向があることが分かった．つまり，成長論文はいわゆる Cutting Edge と言われる最先端の分野の論文である割合が高いといえる．その一方で，「遠くの距離にある論文を引用している」性質は確認できなかった．

4.5 考察

1 次の近接性と 2 次の近接性をもとに Embedding した結果が異なったことは，それぞれの近接性の定義が異なることに起因する．

1 次の近接性を用いた場合，成長論文の特徴量 DC(引用先論文への平均距離) の値が高いことが分かったが，DA(全ノードへの平均距離) は他の論文と大きな差はなかった．1 次の近接性を使用した Embedding(LINE 1st) は直接的なネットワークの接続をもとにしており，分野のクラスタ性を捉えたものになったと考えられる．よって，特徴量 DC によって，「成長論文は自身の分野以外を俯瞰している」と言われる性質を検出できたといえる．

2 次の近接性を用いた場合，成長論文は DA(全ノードへの平均距離) が大きい事がわかったが，DC(引用先論文への平均距離) は他の論文と大きな差がなかった．2 次の近接性は，論文の引用先による間接的な距離をもとにしており古い論文への引用に表現ベクトルが引き寄せられることはない．従って 2 次の近接性を用いて Embedding することで，図 1b,1e のように一方向に論文の領域が成長していることが観察されたといえる．その成長領域の端に位置している論文が Cutting Edge であると考えられる．このようなことから，成長論文は Cutting Edge にある割合が高く，それは DA(全ノードへの平均距離) を用いて定量化し

表 2: 成長論文予測結果

Features	Nanocarbon			Solar Photo			Complex Networks		
	Precision	Recall	F-Value	Precision	Recall	F-Value	Precision	Recall	F-Value
基本特徴量	0.54	0.09	0.15	0.52	0.18	0.27	0.6	0.21	0.31
基本特徴量 + Embedding 特徴量	0.51	0.21	0.3	0.3	0.35	0.32	0.53	0.32	0.4
基本特徴量 + 表現ベクトル	0.53	0.18	0.27	0.53	0.24	0.33	0.58	0.25	0.35
基本特徴量 + Embedding 特徴量 + 表現ベクトル	0.5	0.25	0.33	0.46	0.32	0.38	0.52	0.25	0.34

て理解できることが分かったとえいる．

5. 成長論文予測

以上の結果より定義した特徴量 DA,DC により成長論文の性質を推定することができた．本節では，論文集合の中から成長論文を推定するタスクにその特徴量を応用した．また同時に，表現ベクトルをそのまま用いることも成長論文の推定精度に寄与するかについて確認を行った．

5.1 タスク

論文出版直後のデータより 3 年に被引用数が上位 10%となる論文かどうかを判定することをタスクとする．出版直後のデータセットを用いるため 2014 年末に取得した論文の引用ネットワークより，2014 年度の論文への被引用（2014 年度の論文同士の引用）を除いたものを使用する．

5.2 予測モデル

2010 年までの引用ネットワークをもとに，2010 年度に出版された論文のうち成長論文（2011 年度の被引用数が top10%以内）をの特徴量をもとにロジスティック回帰により分類するモデルを作成した．そのモデルを 2011 年度の論文に適用し 2014 年度の被引用数を予測し，2014 年度の成長論文をどれだけ推定できたかを Precision，Recall，F-value で評価した．

特徴量として，一般的な引用予測に使用される「基本特徴量」，本論文で定義した「Embedding 特徴量」および Embedding で得られた「表現ベクトル」を用いた．それぞれの詳細を以下の表 3 に示す．

表 3: 使用した特徴量

グループ	詳細
基本特徴量	PageRank, 媒介中心性，次数中心性 次数中心性（In），次数中心性（Out） クラスタ係数 引用先論文の上記の MAX 引用先論文の上記の Min 引用先論文の上記の Average
Embedding 特徴量	DA(1st), DA(2nd)，DC(1st)，DC(2nd)
表現ベクトル	Embedding (LINE 1st), Embedding (LINE 2nd)

5.3 予測結果

表 1 のデータセットに対しそれぞれ実験を行った．その結果を表 2 に示す．

表 2 より，定義した Embedding 特徴量および表現ベクトルの両者ともに成長論文の推定に寄与したといえる．いずれのデータセットにおいても，これらを用いることで予測精度は F 値 0.33〜0.38 程となった．Embedding により算出した特量量が寄与しているその理由は，4 節で示したようにそれらの特徴量が成長論文の性質を捉えているからであると考えられる．Embedding により算出された特徴量を用いた場合に，Recall が上昇しているのは，それらの特徴量が，萌芽領域を捉えるという既存特徴量では把握できない論文を把握できているからだと推測される．また，表現ベクトルをそのまま予測に使用する手法も予測結果に寄与していることが分かった．これは，4 節の図 1b,1e のような成長論文が集まっている領域をロジスティック回帰により学習できたからだと考えられる．

6. 結論

本論文では Network Embedding で得られた表現ベクトルが意味を持つことを示すため，論文の引用ネットワークをデータセットとし，特徴量 DA「全ノードへの平均距離」および DC「接続先ノードへの平均距離」を用いて解析を行った．成長論文が Cutting Edge にあることが特徴量 DA を用いた定量的な分析によりわかった。また成長論文は自身の領域から遠い論文を引用していることが特徴量 DC を用いた分析により分かった．さらに，これらの特徴量を使用することで成長論文の予測精度向上に寄与することが分かった．

以上の分析から Network Embedding で算出された表現ベクトルは有用な意味を含んでおり，ラベル推定やリンク予測以外にも応用先があることが分かったといえる．今後はこのような分析を多くのネットワークを対象として行っていくことで，Network Embedding で算出された表現ベクトルをの意味の考察を深めていく必要があると考えられる．

また，本論文で定義した特徴量は引用論文の推定のみな

らず，SNS 上の情報伝播の解析やユーザーの属性推定に適用できると考えられる．具体的には，特徴量 DA はユーザーの SNS 上の役割や関係性の推定に用いることが可能であるし，特徴量 DC はユーザーのコミュニケーションの性質の理解に有用であると考えられる．

謝辞

本研究は NEDO の委託事業「学術産業技術俯瞰システム開発プロジェクト」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] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu 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.

[2] Bryan Perozzi, Rami Al-Rfou, and Steven 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.

[3] Yun Mao and Lawrence K Saul. Modeling distances in large-scale networks by matrix factorization. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 278–287. ACM, 2004.

[4] William Thomas Tutte. How to draw a graph. *Proc. London Math. Soc*, 13(3):743–768, 1963.

[5] Gerard Salton and Michael J McGill. Introduction to modern information retrieval. 1986.

[6] Roger Guimera and Luis A Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[8] Shanshan Zhang, Ce Zhang, Zhao You, Rong Zheng, and Bo Xu. Asynchronous stochastic gradient descent for dnn training. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6660–6663. IEEE, 2013.

[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[10] Han Hee Song, Tae Won Cho, Vacha Dave, Yin Zhang, and Lili Qiu. Scalable proximity estimation and link prediction in online social networks. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 322–335. ACM, 2009.

[11] Dashun Wang, Chaoming Song, and Albert-László Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[12] Tanmoy Chakraborty, Suhansanu Kumar, Pawan Goyal, Niloy Ganguly, and Animesh Mukherjee. Towards a stratified learning approach to predict future citation counts. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 351–360. IEEE Press, 2014.

[13] Feruz Davletov, Ali Selman Aydin, and Ali Cakmak. High impact academic paper prediction using temporal and topological features. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 491–498. ACM, 2014.

[14] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.