The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

# 異種ネットワーク上のノードエンベディング法による萌芽的研究分野特定のための分散表現抽出

The Representation Extraction for Emerging Research Fields Using an Embedding Method for Heterogeneous Networks

大知 正直 *1 Masanao Ochi　城 真範 *2 Masanori Shiro　森 純一郎 *1 Jun'ichiro Mori　浅谷 公威 *1 Kimitaka Asatani　坂田 一郎 *1 Ichiro Sakata

*1東京大学工学系研究科技術経営戦略学専攻
Department of Technology Management for Innovation, Graduate School of Engineering, The University of Tokyo

*2産業技術総合研究所 人間情報研究部門
HIRI, National Institute of Advanced Industrial Science and Technology

It's important to identify promising research early to determine which research to invest. In addition, it's necessary to develop a technology for automatically predicting future research trends because of increasing the number of publication and the research fragmentation. There are many metrics for research performances and it depends on the objective which future trends to show. So, the problem is developing the technology to predict research trends for various metrics. Therefore, in this paper, we propose a method to extract distributed representations for automatically predicting various research metrics in the future using various heterogeneous network information with research papers. Experimental results show that prediction of the reference relation between the research papers was 95.6% F-value, and the *h*-index after three years from publication was 64.4% under certain conditions. The first result shows that the proposed method can sufficiently map the reference relations to a vector space. On the other hand, the prediction accuracy of the future *h*-index is equivalent to the comparison method, and further research needs. The results suggest that distributed representations of heterogeneous networks for scientific paper may be the basis for the automatic prediction of technology trends.

## 1. はじめに

研究への投資戦略策定のために，早い段階で有望な研究や研究分野を特定することは重要である．膨大な知識を俯瞰的に捉え，幅広く将来の技術開発の方向性を評価しようという試みは，技術フォアサイト，ホライズンスキャニング，技術フォアキャスティング，インパクトアセスメント等と呼ばれる．特に最近では，このような活動を専門的に行う政府機関が多くの国々*1 で設立され，政策決定の場でもこれらの試みから得られる知見を活かそうとしている．こうした試みは従来より，専門家に対するアンケートやワークショップによる T-plan 法や Delphi 法，SWOT 分析を用いたもので行われてきた．しかし，近年の論文出版数の急増及び専門知識の細分化によって，少数のメンバーによって学術研究の動向を分析することは困難となっている．また，より広い分野や，分野を横断した研究に対する需要が高まり，研究者個人に依存した技術動向予測は難しくなってきている．このような中で，近年は論文や特許を直接分析し，意思決定に役立てようという試みが盛んに行われている．従来，論文や特許を直接分析する場合は，科学技術の現状を指数化することで・その影響を明らかにしようとすることが多かった．論文そのものでは被引用数，出版した論文誌ではインパクトファクター（IF），研究者個人では *h*-index などが挙げられる．

本稿では，出版社の持つ論文データの様々な情報をネットワーク表現し，将来の技術動向を予測するための分散表現を抽出することを目的とする．これによって，単に論文そのものが出版後どのようなインパクトを与えるかということのみならず，論文の著者や研究グループが近い将来にどのような研究に興味を持ち，どの程度のインパクトを与えるかということが定量評価できる．具体的には，ネットワークから各ノードの分散表現を抽出し，異種ネットワーク向けに改変した新たな手法を提案する．また，本研究では深層型のグラフ分散表現抽出モデルではなく，ランダムウォーク型のグラフ分散表現抽出モデルを採用した．これによって，大規模な異種ネットワークからノード分散表現を抽出することが可能になる．抽出した分散表現を用いることで，グラフデータ全体を用いずとも近傍のノードやコミュニティを検出することが可能になる．

本研究の貢献は以下の３点に要約される．

- 出版社の持つ論文データをネットワーク化し，分散表現を抽出する手法を提案をする．
- 抽出した分散表現によって，リンク予測に利用でき，グラフ構造を十分に抽出できることを示す．
- 抽出した分散表現によって，将来的な研究者の *h*-index の予測の可能性が示唆する．

## 2. 関連研究

これまでの多くの研究は，科学と技術に関する新興市場を予測し，推定するための方法を提案してきた．Fujita らは，論文の引用関係に関する複数のネットワークを作成し，クラスタリングを行い，最新の研究分野の検出の有効性を分析した [2]．Dong らは論文の出版後５年後の著者の *h*-index を予測した．論文のインパクトは６つの要因を用いて定義される，すなわち，著者，内容，出版社，引用，共著者，年代順である．その研究に用いられたデータはコンピュータサイエンスに関する

連絡先: 大知 正直，〒 113-8654 東京大学工学系研究科技術経営戦略学専攻，東京都文京区本郷７－３－１，"masanao.oochi@gmail.com"

*1 例えば，欧州議会科学技術選択評価委員会 (STOA) 内に設立されたユニット等 http://www.europarl.europa.eu/RegData/etudes/IDAN/2015/527415/EPRS_IDA%282015%29527415_REV1_EN.pdf

200 万件の論文情報である [1]．また，Sasaki らは出版直後のネットワーク的な情報のみで，その後の論文の被引用数の予測を行った [5]．このように研究分野の将来的な発展の可能性を早期に発見しようという試みは盛んに行われており，本研究もその 1 つである．

一方，2015 年からグラフを直接ベクトル空間に写像しようという試みが行われている [4, 7]．この試みは様々に発展し現在も盛んに研究されている．深層化しようという代表的な試みは GCN[3] と呼ばれるものである．しかし，現状ではグラフの規模も 10 万ノード程度でニューラルネットワークの階層も 2 層ほどで成果を報告しているものが多く，大規模化，深層化にはなお多くの課題が残されている．本研究では，比較的大規模化しやすいランダムウォークベースのグラフ分散表現の抽出手法を採用し，それを論文データ上で構築する異種ネットワーク向けに拡張したものを提案する．テキストデータ向けの異種ネットワーク分散表現獲得モデルがすでに報告されており [6]，本稿では，この手法を論文データ用に拡張し，萌芽的研究分野特定に有効か評価を行う．

## 3. 提案手法

### 3.1 論文データを用いた異種ネットワーク

本稿では論文データを用いた異種ネットワークを用いる．用いる異種ネットワークの概略を図 1 に示す．解析対象とするネットワークは 5 つある．そして，それぞれのグラフは，一部のノードを他のグラフと共有しており，全く共有していないグラフは存在しない．まず，citation network で，これは論文の持つ引用関係をグラフにしたものである．次に，paper-author network は，論文と著者間をエッジで結んだものである．ただし，共著者間でエッジを結ばず，必ず論文とその著者で結ぶものとする．author-institute network は著者と著者の所属する研究機関の所属関係をエッジで結んだものである．また，論文の特徴を表すために設定されたキーワードを論文と結んだものを paper-keyword network とする．最後に論文が出版された雑誌と結びつけた paper-journal network を加える．

これらそれぞれのグラフのほとんどは，論文だけ，著者だけ，といった単一種類のノードのみで構成されたグラフではなく，異なる種類のノード群で構成された複数の異種ネットワークである．

### 3.2 提案手法の概要

本節では，まず論文データから抽出した異種ネットワークの各ノードを単一のベクトル表現空間へ写像する方法について説明する．次に，萌芽的研究分野を特定するために有効な様々なタスクへと適用する手法を説明する．提案手法の概要を図 2 に示す．

#### 3.2.1 異種ネットワーク上のノードエンベディング法

本稿で用いる複数の異種ネットワーク上のノードを同一のベクトル表現空間へ写像を行う手法の説明を行う．まず，複数の異種ネットワークの集合 $\boldsymbol{G}$ を以下のように表す．ここで,$G$ はグラフ，$V$ はノード集合，$E$ はエッジ集合とする．

$$\boldsymbol{G} := \left\{ G^l = (V^l, E^l) \middle| 1 \leq l \leq |\boldsymbol{G}| \right\} \tag{1}$$

ここで，各 $G^l$ は図 1 に示した author–institute network や citation network のような複数の異種ネットワークで，$|\boldsymbol{G}|$ は対象とする異種ネットワークの数である．また，それぞれのグラフはノードとエッジで構成された無向グラフ $G^l := (V^l, E^l)$ とし，それぞれのグラフで共有するエッジは無いが，ノードは一部隣接するグラフと共有されているものとする．このような異種ネットワーク $\boldsymbol{G}$ 上のあるノード $v_i \in \boldsymbol{V}$ をベクトル空間のあるベクトル $\overrightarrow{u_i} \in \boldsymbol{U}$ へと写像する方法を考える．ただし，$\boldsymbol{V}$ は各異種ネットワーク上のノード群 $V^l$ 全てを指す．

表 1: 学習アルゴリズム.

学習アルゴリズム.

**Input**: $\boldsymbol{G}, T, \rho_0, K, D$.
**Output**: $\boldsymbol{U}$.
各ノードの写像ベクトル，コンテクストベクトル $\boldsymbol{U}, \boldsymbol{U}'$ を $D$ 次元で初期化.
**for** $t = 1$ **to** $T$
  $\rho = \rho_0(1 - \frac{t}{T})$
  **for** $l = 1$ **to** $|\boldsymbol{G}|$
    エッジ $e_{ij}^l$ を $G^l$ からサンプリング.
    ノード $v_i^l, v_j^l$ に対応する写像ベクトル，コンテクストベクトルを $\boldsymbol{U}, \boldsymbol{U}'$ から読み出す.
    写像ベクトルの更新: $\overrightarrow{u^{t+1}} = \overrightarrow{u^t} - \frac{\rho_t}{w^l} \frac{\partial O}{\partial \overrightarrow{u}}$
    コンテクストベクトルの更新: $\overrightarrow{u'^{t+1}} = \overrightarrow{u'^t} - \frac{\rho_t}{w^l} \frac{\partial O}{\partial \overrightarrow{u'}}$
END

まず，ある特定のグラフ上で隣接するノード $v_j$ へのエッジの重み $w_{ji} = P(v_j|v_i)$ を写像したベクトル表現で以下の式によって算出できるとする．簡略化のため，特定のグラフを対象にし，添字 $l$ を省略した．

$$\hat{P}(v_j|v_i) = \frac{\exp(\overrightarrow{u_j'}^{\mathsf{T}} \cdot \overrightarrow{u_i})}{\sum_{k \in |V|} \exp(\overrightarrow{u_k'}^{\mathsf{T}} \cdot \overrightarrow{u_i})} \tag{2}$$

ここで $|V|$ はグラフ上のノード数とし，$\overrightarrow{u_i}$ をノード写像ベクトル，$\overrightarrow{u_j'}$ をノード $v_i$ からのコンテキストベクトルとする．このコンテキストベクトルは計算上用いるのみで，本稿ではノード写像ベクトルを求めることに興味がある．そして，この式を用いることで，元のグラフ上でのエッジの重みとの差を定式化できる．$P(\cdot|v_i)$ をノード $v_i$ からすべてのノードへのエッジの重み（エッジが存在しないノードへは重み 0 で接続されているとする）の分布とすると，ベクトル表現から算出したエッジの重みの分布との距離の和 $O$ を $\sum_{i=1}^{|V|} \lambda_i d\left(P(\cdot|v_i), \hat{P}(\cdot|v_i)\right)$ によって計算できる．例えば距離関数 $d$ に KL 擬距離を採用し，係数 $\lambda_i = \sum_{j=1}^{|V|} w_{ji}$ としたとき，推定する $\hat{P}(\cdot|v_i)$ の変数部分のみに注目すると $O$ は以下のように近似できる．

$$O \approx -\sum_{(i,j) \in E} w_{ji} \log \hat{P}(v_j|v_i) \tag{3}$$

この特定のグラフ $G^l$ 内のノード集合 $V^l$ をベクトル空間に写像したベクトル集合 $\boldsymbol{U^l} := \left\{ \vec{u_i^l} \middle| 1 \leq i \leq |V^l| \right\}$ との距離の和 $O^l$ を異種ネットワーク $\boldsymbol{G}$ 内のすべてのグラフに対して和をとる．

$$O = \sum_{l=1}^{|\boldsymbol{G}|} O^l \tag{4}$$

この式 4 の $O$ を最小化する．

#### 3.2.2 学習の手順

3.2.1 節で説明した更新式を以下のアルゴリズムによって順次更新し，ノード写像ベクトルの最適化を行う．アルゴリズム内の $T$ は学習回数，$K$ は負例サンプリングの回数，$D$ は写像するベクトル空間の次元数を表す．

#### 3.2.3 萌芽的研究分野特定の手順

本研究では，萌芽的研究分野特定を行うことを可能にする分散表現抽出を目的とする．その初段として，論文の将来的な被

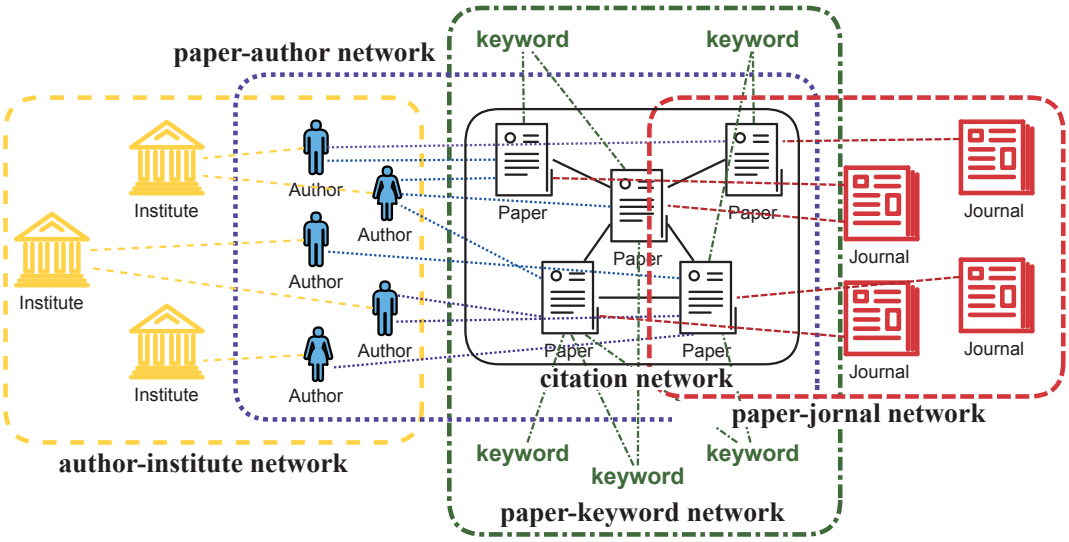


図 1: 本研究で用いる異種ネットワークの概略図.

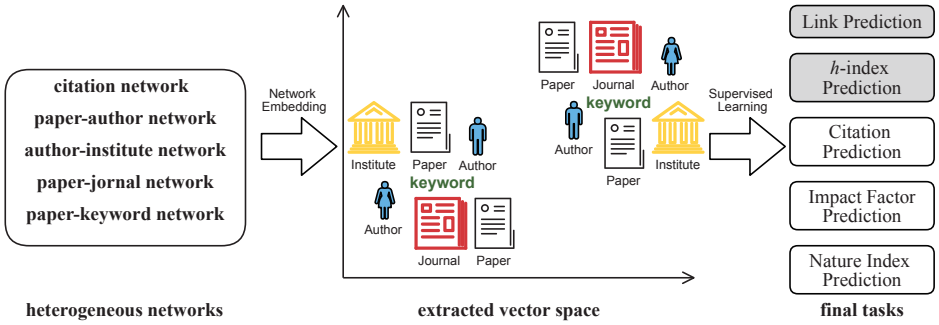


図 2: 本稿で提案する手法の概要．final tasks 部分で色付きの Link Prediction, $h$-index Prediction が今回試行したタスク

引用数と研究者の将来的な $h$-index がトップ $x$%に入るか検証を行う．これによって，異種ネットワークから抽出した分散表現が将来の研究動向の予測において，有用であることを示す．具体的には，ある年 $(Y)$ から $n$ 年後のある対象 $(ID)$ の持つ何らかの指標 $I_{ID,Y+n}$ の予測を以下のロジスティック回帰を用いて行う．

$$\hat{I}_{ID,Y+n} = \sigma\left(\vec{w}^{\mathsf{T}} \cdot (\boldsymbol{U}_s^{\mathsf{T}} \cdot \vec{u}_{ID,Y})\right) \tag{5}$$

この式で，$\vec{w}$ は最適化する重みベクトル，$\boldsymbol{U}_s$ は適当に対象をサンプルした対象群 $\boldsymbol{s}$ のベクトルを並べた行列，$\vec{u}_{ID,Y}$ はある対象のある年の異種ネットワークで学習したベクトルである．これによって，ある対象 ID とサンプルした対象群との距離を特徴量として，その重み $\vec{w}$ を学習することによって，指標 $I$ がトップ $x$%にはいるかどうかの予測を行う．また，$\vec{w}$ は，Y よりも過去 ($m$ 年前) の時点で学習しておく必要があるため，訓練時には $Y-m$ 時点で，$Y-m+n$ 年での指標 $I_{Y-m+n}$ の予測を行い学習する．そこで最適化を行った $\vec{w}$ を指標 $I_{Y+n}$ の予測に用い，評価を行う．

## 4. 実験と結果

### 4.1 使用するデータ

使用するデータは国際的に多数の学術雑誌を発行している Elsevier より提供を受けた”(TITLE-ABS-KEY(nano AND carbon) OR TITLE-ABS-KEY(gan) OR TITLE-ABS-KEY(solar AND cell) OR TITLE-ABS-KEY(complex AND networks)) AND PUBYEAR AFT 2006”というクエリに Scopus 上でヒットする文献データセットを用いる．このデータセットは，Nano Carbon, GaN, complex networks, Solar Cell に関する学術文献に関するデータで，342,785 件の論文データを含んでいる．論文データごとに，著者，研究機関，引用文献，雑誌名，アブストラクトの情報を持っている．

表 2: データセット内の $h$-index の順位.

| Ranking @2016 | Name | $h$-index 2009 | 2013 | 2016 |
|---|---|---|---|---|
| 1 | Michael Gratzel | 23 | 72 | 116 |
| 2 | Mohammad K.haja Nazeeruddin | 13 | 45 | 80 |
| 3 | Anders Hagfeldt | 13 | 50 | 74 |
| 4 | Shaik M.ohammed Zakeeruddin | 15 | 47 | 64 |
| 5 | Henry J. Snaith | 7 | 29 | 63 |
| 6 | Li Cheng Sun | 13 | 44 | 60 |
| 7 | Yong-fang Li | 7 | 34 | 57 |
| 8 | Christoph J. Brabec | 12 | 33 | 57 |
| 9 | Alan J. Heeger | 8 | 32 | 55 |
| 10 | Frederik Christian Krebs | 12 | 41 | 55 |

### 4.2 リンク予測

#### 4.2.1 実験条件

本節では抽出した分散表現がグラフ構造を十分に学習できているか評価する．グラフ上からランダムに選択したノードのペア間にリンクが存在するかどうかを学習し，テストデータで十分に予測できるか実験する．リンクの存在の有無は $\hat{e}_{ij} = \sigma\left(w\vec{u}_i^{\mathsf{T}} \cdot \vec{u}_j\right)$ で判定を行う．$\hat{e}ij >= p$-th の場合，その 2 つ

のノード間にはエッジが存在していると判定する．訓練のためにグラフから正例としてエッジのあるノードのペア，負例としてエッジの無いノードのペアを同数ずつ合計で 4358 万件サンプルする．テストデータ用に訓練データの数の 10%を同数ずつサンプルする．訓練データで最適化を行った $w$ を用いて，テストデータのペア間のエッジの有無を予測し，その精度を評価する．比較手法として，ノード間のエッジの有無を 0.5 でランダムで選択する手法を用いる．

#### 4.2.2 実験結果

実験結果を表 3 に示す．リンク予測に関しては総合的に見て高い精度で予測できており，本手法が十分にグラフの構造を学習できていることを示している．F1-value では $p$-th= 0.50 の場合に, 提案手法の結果が 0.956 となっており, 比較手法と比較して 0.458 ポイント高い.

### 4.3 $h$-index 指標予測

#### 4.3.1 データセット内の $h$-index 指標

$h$-index 指数を”solar cell”というクエリで取得した scopus のデータセット内で計算した結果を表 2 に示す．1 位の Michael Graätzel は，色素増感型太陽電池の発明者で，2016 年時点で，色素増感太陽電池での最高記録となる 15%のエネルギー変換効率を達成しており，2019 年 1 月現在での Google Sholar 上の $h$-index は 253 である．その他の研究者についても概ね大きな $h$-index を示しており，太陽電池に関する研究者の $h$-index の順位は実際の順位と概ね一致している．

#### 4.3.2 実験条件

本稿では，$Y = 2013, n = 3, m = 4$ とし，サンプルした対象数 $|s| = 20$ とし，指標のトップ $x = 20\%$ に入るかどうかの予測を行う．つまり，2013 年の異種ネットワークを用いて，分散表現を抽出し，2016 年の著者の $h$-index がトップ 20%に入るかどうかの予測を行う．予測にあたって，学習には 2009 年の異種ネットワークから抽出した分散表現と 20 人サンプルした分散表現とのそれぞれの距離を特徴量として，2012 年時点での $h$-index の予測し，実際の値との最適化を行う．そして学習した $\vec{w}$ を用いて 2016 年の予測を行い，結果の評価を行う．また今回は比較手法として，$Y$ 時点での $h$-index をそのまま $Y + n$ の結果として用いる．

#### 4.3.3 実験結果

実験結果を表 4 に示す．表で，提案手法はロジスティック回帰による出力を行うので，その値のしきい値として $p$-th を設定している．このしきい値による違いと比較手法との差について確認する．まず，Precision は $p$-th= 0.80 のときに 0.992 で，比較手法の 0.704 と比較し，0.288 ポイント向上している．また，Recall は $p$-th= 0.06 のときに 0.717 で，比較手法の 0.652 と比較し，0.065 ポイント向上している．しかし，一方で F1-value では，提案手法は最大でも 0.644 であり，比較手法の 0.677 と比較し，0.033 ポイント下回った．

表 3: リンク予測の結果.

| Method | $p$-th | Precision | Recall | F1-value |
|---|---|---|---|---|
| proposed | 0.25 | 0.912 | **0.992** | 0.950 |
| proposed | 0.50 | 0.943 | 0.971 | **0.956** |
| proposed | 0.75 | **0.965** | 0.911 | 0.937 |
| baseline | | 0.501 | 0.495 | 0.498 |

表 4: 将来の $h$-index 予測の結果.

| Method | $p$-th | Precision | Recall | F1-value |
|---|---|---|---|---|
| proposed | 0.06 | 0.585 | **0.717** | 0.644 |
| proposed | 0.80 | **0.992** | 0.0606 | 0.114 |
| baseline | | 0.704 | 0.652 | **0.677** |

## 5. 考察

まずリンク予測については，提案したアルゴリズムによって十分にグラフ構造を学習できていることが確認できた．特にノード間のエッジの有無について 95%以上の精度で予測できており，グラフ上で隣接するノード群を十分近傍にベクトル空間上へ写像できていると考えられる．一方で，将来の $h$-index 予測では，個別の Precision，Recall といった指標は $p$-th の値を調整することで，高い値を得ることができる．しかし，両者の調和平均を取った F1-value においては最高で 0.644 を示しているものの比較手法をわずかに下回っている．パラメータを変化させることでこの精度が大きく変わることがないか今後検討する必要がある．

## 6. 結論

本稿では，萌芽的研究分野特定のための分散表現抽出手法を提案した．提案手法では，グラフのノード間のエッジの存在の有無を 95%以上の精度で予測でき，グラフ構造を十分にベクトル空間に写像できていることがわかった．一方で，将来の $h$-index 予測では precision，recall それぞれにおいては高い精度を示す結果を得ているが，F1-value においては比較手法を下回る結果であった．この結果は，抽出した分散表現によって，将来的な研究者の $h$-index の予測の可能性が示唆する．今後は，提案手法のパラメータの敏感性を調査したり，さまざまな指標予測を行ってみることで，本手法の限界，安定性について比較を行う予定である．

## 参考文献

[1] Yuxiao Dong, Reid A. Johnson, and Nitesh V. Chawla. Will this paper increase your h-index?: Scientific impact prediction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 149–158, New York, NY, USA, 2015. ACM.

[2] Katsuhide Fujita, Yuya Kajikawa, Junichiro Mori, and Ichiro Sakata. Detecting research fronts using different types of weighted citation networks. *Journal of Engineering and Technology Management*, 32:129 – 146, 2014. Special Issue on Emergence of Technologies: Methods and Tools for Management.

[3] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[4] 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.

[5] Hajime Sasaki, T Hara, and Ichiro Sakata. Identifying emerging research related to solar cells field using a machine learning approach. *Journal of Sustainable Development of Energy, Water and Environment Systems*, 4:418–429, 12 2016.

[6] Jian Tang, Meng Qu, and Qiaozhu 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*, KDD '15, pages 1165–1174, New York, NY, USA, 2015. ACM.

[7] 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*, WWW '15, pages 1067–1077, New York, NY, USA, 2015. ACM.