DEIM Forum 2021 F33-1

キャリアパスを考慮したネクスト企業推薦手法

福知 侑也† 馬 強†

† 京都大学大学院情報学研究科 〒 606–8501 京都府京都市左京区吉田本町 36-1
E-mail: †fukuchi@db.soc.i.kyoto-u.ac.jp, ††qiang@i.kyoto-u.ac.jp

あらまし 近年オンラインの採用プラットフォームにて転職活動を行う機会が多くなっており，企業や求職者の特徴を考慮した職業推薦の手法が多く提案されている．しかしながら，既存手法の多くは求職者の職歴の系列性や順序，いわゆるキャリアパスを十分に考慮していない．そこで本研究では，企業と求職者からなるヘテロネットワークを構築し，企業間やユーザ間の関係を分析し，グラフエンベッディングと LSTM モデルを用いて求職者のキャリアパスを考慮したネクスト職業推薦手法を提案する．実験の結果 HitRate と MRR の 2 つの指標において，提案手法がベースラインを上回ったことを確認した．

キーワード 情報推薦， グラフ， 職業， シーケンシャル推薦， 機械学習

1 はじめに

近年オンラインの採用プラットフォームにて就職活動 (転職活動) を行う機会が多くなっている．例えば有名なビジネス SNS であるリンクトインでは約 7 億人の登録者，約 2000 万件の求人情報の掲載がされている [1]．また近年，従業員が頻繁に転職することが当たり前になっている．2006 年から 2010 年の間に卒業した学生は最初の 5 年で平均 2.85 個の職についており，この数字は 1986 年から 1990 年に卒業した学生の 2 倍であると報告されている [2]．求職者は膨大な情報の中から自分にあった企業を探す必要があり，その頻度も増加している．そのため採用プラットフォーム上で適切な企業を求職者に推薦することで，求職者の職業の検索を支援することができる．

ネクスト職業推薦 (もしくはネクスト企業推薦) の既存手法として，ユーザのプロフィール情報または履歴書の情報から特徴量を抽出し，それらの特徴量を用いて機械学習モデルを学習し，ネクスト職業の推薦を行う手法が多く提案されている [3] [4] [5]．協調フィルタリングを用いた手法 [3]，ナイーブベイズを用いた手法 [4] や GBDT(Gradient Boosting Decision Tree) モデルを用いた手法 [5] など多岐にわたる．しかしながら既存手法は求職者の職歴の系列性や順序，いわゆるキャリアパスを十分に考慮していない．キャリアパスには求職者のキャリアの選択という過去の意思決定の情報が含まれており，これらの系列の情報は非常に重要であると考える．また直近のキャリアが昔のキャリアに比べて比較的重要になるため，現職までの全てのキャリアを異なる重みで用いることが望ましい．

採用市場においてある企業からある企業への人材の流入出の性質を捉えることも企業の推薦において重要である．多くの人は現職と似たような企業，もしくは現職と関連がある企業を転職先として求める．その性質は採用市場の企業間の人材の流れに反映されていると考える．Zhang [6] は企業間の人材の流れに着目し，タレントフローネットワークを作成することで企業の競合分析手法を提案している．

そこで本研究では，人材の流れに着目した企業間の関係性と，企業とユーザの関係性をヘテログラフで表現し，キャリアパスを考慮したネクスト企業推薦手法を提案する．はじめにユーザのプロフィールデータから企業ノードとユーザノードからなるヘテログラフを構築する．このグラフの企業ノード間のエッジは人材の流れを表し，企業ノードとユーザノードの間のエッジはそのユーザがその企業に所属していることを表す．次に Metapth2vec [7] というグラフエンベッディング手法を用いて，企業間の関係や企業とユーザ間の関係を考慮したベクトル表現を得る．その後 LSTM モデル [8] を用いてユーザのキャリアパスを考慮したネクスト企業推薦モデルを構築する．

本稿の構成は以下の通りである．第 2 節では関連する研究について説明をする．第 3 節では提案手法についてその順序とともに説明する．第 4 節では実験に用いるデータセットや実験の結果について示す．最後に第 5 節でまとめる．

2 関連研究

Zhang [3] らは協調フィルタリングを用いた学生への職業推薦手法を提案している．はじめに学生の履歴書データを用いて学生と職業からなるブール値行列を作成する．ブール値行列の値は過去に学生がその職業に応募したかどうかを表す．その行列にユーザベースとアイテムベースの協調フィルタリングそれぞれを適用し，職業推薦を行う．また，Zhang らは Recall, Precision，F-Score といった評価指標においてユーザベースの協調フィルタリングとアイテムベースの協調フィルタリングの比較を行い，アイテムベースの協調フィルタリングがより高い精度を出したことを報告している．

Paparrizos ら [4] はナイーブベイズモデルを用いてネクスト職業の推薦を行っている．実験に使用するデータはある採用プラットフォーム上の公開プロフィールデータを用いる．各ユーザごとにキャリアパスを (前の職業，次の職業) というペアに分解し学習データとしている．また図 1 に示す三つの学習データセットを作成している．I はデータセット内の出現頻度上位

Setup	Data sample	Set sizes	
		Train set	Test set
I	Top 100 universities + top 100 companies	65,622	28,124
II	Top 100 companies	52,142	22,346
III	Top 25 companies	45,891	19,668

図 1　実験に用いた三つのデータセット (Paparrizos 2011 Table3 [4])

100 の大学と企業間の遷移のみを用いている．IIはデータセット内の出現頻度上位 100 の企業間の遷移のみを用いている．IIIはデータセット内の出現頻度上位 25 企業間の遷移のみを用いている．そして最後に 3 つのそれぞれのデータセットに対して学習と予測を行い，一番出現頻度の多い企業を推薦した場合をベースラインとして比較を行っている．Paparrizos らは提案手法がベースラインの評価を大きく上回ったことを報告している．

Snorre [5] は GBDT(Gradient-Boosted Decision Trees) モデルを用いて，ユーザ情報を特徴量としてネクスト職業の推薦を行っている．この研究では，2012 年から 2016 年のデンマーク人の背景情報を含んだ採用市場データを用いている．特徴量として用いるユーザ情報には職歴，性別，年齢，居住地などが含まれる．GBDT モデルには XGBoost [9] を採用している． XGBoost モデルを用いたネクスト職業推薦は MRR@10 と Recall@10 の二つの評価指標において，アイテムベースの協調フィルタリングを用いた場合と最も人気な職業を推薦した場合よりも精度が高かった．また、特徴量重要度を計算し，前職の企業の特徴量が 1 番重要であることが分かった．上記で紹介した 3 つの関連研究において，使用している機械学習モデルは異なるものの，入力と出力の機構は全て現職を入力としネクスト職業を出力する機構である．つまりキャリアパスを十分に考慮していない．キャリアパスには求職者のキャリアの選択という過去の意思決定の情報が含まれており，これらの系列の情報は重要であると考える．そこで本研究では過去のキャリアパスを入力とし，ネクスト職業を出力とする機構を持つシステムの提案を行う．

3　提 案 手 法

本節では, 本論文で提案するグラフの生成方法とベクトルの表現の獲得方法, および学習方法を説明する．提案システムの全体の概観を図 3 に示す．はじめにユーザのプロフィール情報から企業のノードとユーザノードからなるヘテログラフを生成する．次に，グラフエンベッディング手法である Metapath2vec [7] を用いて企業ノードとユーザノードの各ノードのベクトル表現を獲得する．その後ユーザのキャリアパスとそれに対応する企業の各ベクトルを用いて各キャリアパスをベクトルの系列で表すことで学習に用いるデータセットを形成する．最後に，LSTM [8] ベースのモデルを学習し，キャリアのベクトルの系列からネクスト企業の推薦を行う．

3.1　グラフの生成

本研究ではユーザのプロフィール情報から企業ノードとユーザノードの 2 種類のノードを持つ企業-ユーザヘテログラフを生成する．グラフの概観について図 2 に示す．図 2 中で青色で示したノードが企業ノード，紫色で示したノードがユーザノードを表す．企業ノードと企業ノード間の無向エッジは企業間の人材の遷移を表しており，重みには過去にその 2 企業間で転職があったユーザの人数を用いている．一方企業ノードとユーザノード間の無向エッジは過去にそのユーザがその企業に属していたことを表しており，重みは全て 1 とする．またユーザノード間にはエッジをもたない．

以下では企業集合を $\mathbf{C}$, ユーザ集合を $\mathbf{U}$ として企業ノードを $c_i (i \in \mathbf{C})$, ユーザノードを $u_j (j \in \mathbf{U})$ と表記する．また企業とユーザ全体の集合を $\mathbf{V}$ とする．

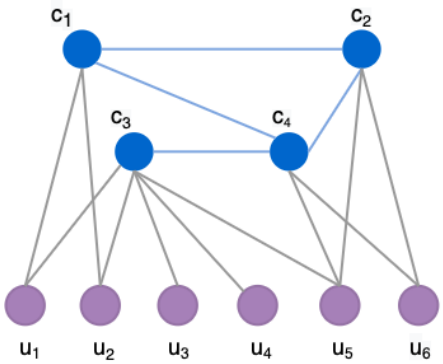

図 2　企業-ユーザのヘテログラフの例

3.2　Metapath2vec を用いたベクトル表現の獲得

Metapath2vec [7] はヘテログラフに対してベクトル表現を獲得することのできる手法である．Metapath2vec において，はじめにメタパス構造を決定する．メタパス構造とは $P : V_1 \xrightarrow{R_1} V_2 \xrightarrow{R_2} \cdots \xrightarrow{R_{l-1}} V_l$ で定義され，$R = R_1 \circ R_2 \circ \cdots R_{l-1}$ はノード V_1 と V_l の意味的な関係を表す．V の添字はノードの種類を表しており，本研究においてはノードの種類は「企業」「ユーザ」の 2 種類である．本実験では企業 - ユーザ - 企業, ユーザ - 企業 - ユーザの 2 つのメタパスを使用する．

次にこのメタパス構造に基づいてランダムウォークを行う．ステップ i における遷移確率は以下のように求める．

$$p(v^{i+1}|v^i, P) = \begin{cases} \frac{1}{|N_{t+1}(v_t^i)|} & (v^{i+1}, v_t) \in E, \phi(v^{i+1}) = t+1 \\ 0 & (v^{i+1}, v_t) \in E, \phi(v^{i+1}) \neq t+1 \\ 0 & (v^{i+1}, v_t) \notin E \end{cases} \tag{1}$$

ここで $v_t^i \in V_t$ であり，$N_{t+1}(v_t^i)$ は v_t^i から距離が 1 であるノードの種類が V_{t+1} と同じであるノードの集合を示す．ランダムウォークによりノードの系列の集合を獲得した後，Skip-Gram によって学習を行いベクトル表現を獲得する．Metapath2vec の具体的なアルゴリズムを Algorithm1 に示す．

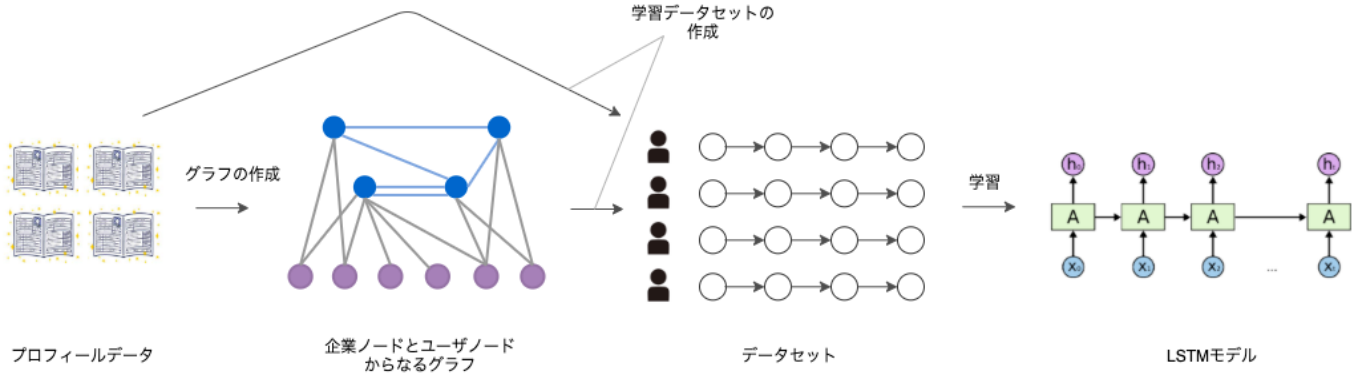

図 3　企業推薦システム全体の概要

Algorithm 1 Metapath2vec algorithm(Yuxiao,2017, ALGORITHM1 [7])

Input: The heterogeneous graph $G = (V, E, T)$, a meta-path scheme P, walks per node w, walk length l, embedding dimension d, Skip-Gram window size k

Output: The latent node embeddings $\mathbf{X} \in \mathbb{R}^{|V| \times d}$

Initialize $\mathbf{X}$;

for $i = 1$ to w **do**
  for $v \in (V)$ **do**
    MP =**MetaPathRandomWalk**(G, P, v, l)
    X =**HeterogeneousSkipGram**(X, k, MP)
  end for
end for
return X;

MetaPathRandomWalk(G, P, v, l)
$MP[1] = v$;
for $i = 1$ **to** $l - 1$ **do**
  draw u according to Eq. 1;
  $MP[i+1] = u$;
end for
return MP;

HeterogeneousSkipGram(X, k, MP)
for $i = 1$ **to** l **do**
  $v = MP[i]$;
  for $j = max(0, i-k)$ **to** $min(i+k, l) \& j \neq i$ **do**
    $c_t = MP[j]$;
    $X^{new} = X^{old} - \eta \cdot \frac{\partial O(X)}{\partial X}$;
  end for
end for

3.3　データセットの作成

LSTM モデルを学習するためのデータセットの作成方法について説明する．はじめに，データセットから各ユーザのキャリアパスを時系列順に並べて取り出す．次に系列の長さによって場合分けし，データセットを作成する．ここで LSTM のウインドウサイズを 3 とし例を挙げて説明する．また，説明変数となるベクトルの系列の集合を $\mathbf{X}$, 目的変数の集合を $\mathbf{Y}$ とする．以下の 3 人のユーザについて考える．

ユーザ 1: $c_1 \to c_2 \to c_3 \to c_4$
ユーザ 2: $c_1 \to c_2 \to c_3$
ユーザ 3: $c_1 \to c_2 \to c_3 \to c_4 \to c_5 \to c_6$

ユーザ 1 からは $\mathbf{x} = \{c_1, c_2, c_3\} \in \mathbf{X}, \mathbf{y} = (onehot(c_4))$ となる $\mathbf{x}, \mathbf{y}$ の 1 つの組を作成する．

ユーザ 2 からは $\mathbf{x} = \{\mathbf{0}, c_1, c_2\} \in \mathbf{X}, \mathbf{y} = onehot(c_3)$ となる $\mathbf{x}, \mathbf{y}$ の 1 つの組を作成する．LSTM の入力に用いるベクトルは固定長であるため長さが不足する分は要素が全て 0 であるベクトルで埋めることで対応する．

ユーザ 3 からは $\mathbf{x_1} = \{c_1, c_2, c_3\} \in \mathbf{X}, \mathbf{y_1} = onehot(c_4)$, $\mathbf{x_2} = \{c_2, c_3, c_4\} \in \mathbf{X}, \mathbf{y_2} = onehot(c_5)$, $\mathbf{x_3} = \{c_3, c_4, c_5\} \in \mathbf{X}, \mathbf{y_3} = onehot(c_5)$ となる $(\mathbf{x_1}, \mathbf{y_1}), (\mathbf{x_2}, \mathbf{y_2}), (\mathbf{x_3}, \mathbf{y_3})$ の 3 つの組を作成する．またユーザのキャリアパスが 1 の場合はデータセットとして用いない．

3.4　LSTM によるネクストジョブ推薦

時系列性を持つデータを学習するニューラルネットワークモデルは複数提案されている．その中でも LSTM モデルは長期情報と短期情報を共に保持し，より精度よく系列データを予測できるという特徴がある [8]．本研究では現職だけでなく過去の職業履歴も考慮してネクスト企業の推薦を行うということが目的であり，LSTM モデルの性質は本研究に適していると考える．よって本研究では LSTM モデルを推薦モデルとして選択する．LSTM モデルの繰り返されるモジュールの機構を図 4 に示した．LSTM モデルの 1 つのモジュールにおける計算の流れは以下の通りである．t 番目の LSTM への入力値を x_t，記憶セルの状態を C_t，出力値を h_t とする．はじめに忘却ゲートの値 f_t を計算する．

$$f_t = \sigma(W_f \cdot [h_{t-1},\ x_t] + b_f) \tag{2}$$

ここで，W_f は f のための重み，b_f は f のための切片，$\sigma(\cdot)$ はシグモイド関数を表す．次に入力調整ゲートの値 i_t と記憶セルの状態の候補値 $\tilde{C}_t$ を計算する．

$$i_t = \sigma(W_i \cdot [h_{t-1},\ x_t] + b_i) \tag{3}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1},\ x_t] + b_C) \tag{4}$$

これらの値を用いて $\tilde{C}_t$ を計算する．

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C_{t-1}} \tag{5}$$

最後に出力ゲートの値 o_t と潜在状態ベクトル h_t を計算する．

$$o_t = \sigma(W_o \cdot [h_{t-1},\ x_t] + b_o) \tag{6}$$

$$h_t = o_t * \tanh(C_t) \tag{7}$$

LSTM モデルの入力は前セクションで示した $\mathbf{X}$ であり，目的変数は $\mathbf{Y}$ である．

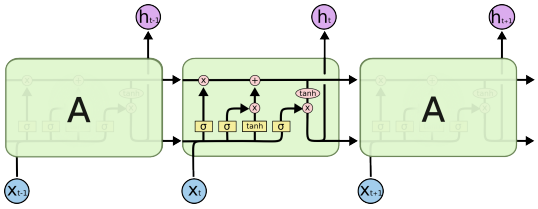

図 4　LSTM における繰り返しモジュール機構 [10]

4　実　　験

4.1　データセット

データセットには世界最大規模のビジネス SNS である Linkedin [11] からユーザのプロフィールを集めた公開データを使用する．データの形式の例を表 4.1 に示す．データセットにはその他のカラムも含まれるが，今回はユーザ ID, 企業名, 始業開始時期のみを使用する．このデータセットは 39,535 行からなり，総企業数は 13,755，総ユーザ数は 6,853 であった．また企業のデータセット内の出現回数は平均が 2.87，最大値が 683，最小値が 1 であり，多くの企業が出現回数が 1 回であることがわかる．

表 1　データセットに含まれる各カラムの値の例

ユーザ ID	企業名	始業開始時期
u1	c1	2010-1-1
u1	c2	2011-1-1
u1	c3	2012-1-1
u2	c2	2010-10-10
u2	c4	2015-1-1
u3	c2	2013-4-5

作成した企業ノードと企業ノード間のエッジのみからなるグラフ (以下企業グラフと呼ぶ) において，企業ノードの次数の分布を図 5 に示す．横軸は次数であり，縦軸はその範囲の次数を持つ企業ノードの数を表している．この図から企業グラフはスケールフリー性を持つグラフであることがわかる．また，企業グラフのノード数は 13,755，エッジの数は 22,072，平均次数は約 3.2 であった．

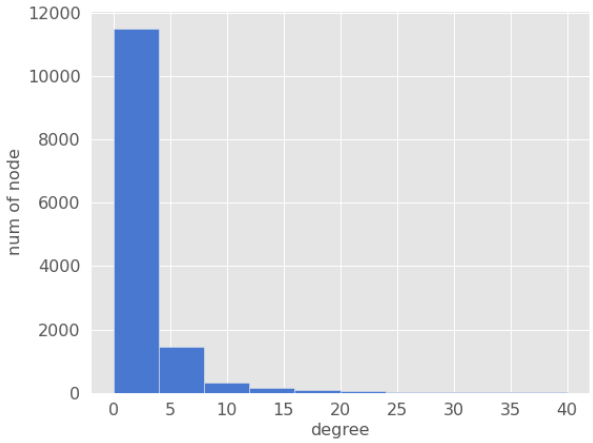

図 5　企業グラフの次数の分布

4.2　評 価 指 標

評価指標には HitRate と MRR(Mean Reciprocal Rank) を用いた．HitRate はデータセットの全レコード数を n, 各レコードを q, 各企業を c とした時，LSTM の出力によるスコア $s(q,c)$ が高いものから順に K 件取得した集合を $I_{q,c}$ とすると

$$HR@K = \frac{1}{n}\sum_{q} |I_{q,K}| \tag{8}$$

と表される．今回は $K = 10, 20, 30, 40, 50$ を用いた．MRR は上位 K 件のスコアを持つ企業において，正解の企業のランキングの逆数の平均として表される. 式は次の通りである．

$$MRR@K = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{rank_i} \tag{9}$$

ここで $rank_i$ は i 番目のレコードの予測スコア上位 K 件におけるランキング位置を表す．また，上位 K 件に正解の企業が含まれない場合は $\frac{1}{rank_i} = 0$ とする．

4.3　実 験 結 果

今回はデータセット内において出現頻度上位 K 位以内の企業を常に推薦する手法をベースラインとし HitRate と MRR の二つの評価指標において提案手法と比較を行う．評価は Fold 数を 5 とした交差検証によって行う．また，提案手法において使用したパラメータを表 4.3 に示す．

表 2　実験に使用したパラメータ

パラメータ名	値
Metapath2vec で学習するベクトルの次元	128
Metapath2vec における各ノードを始点とするランダムウォークの試行数	5
Metapath2vec におけるランダムウォークの長さ	50
Metapath2vec における Skipgram のイテレーション数	5
Metapath2vec における Skipgram のウィンドウサイズ	5
LSTM における学習エポック数	50
LSTM に入力する系列のサイズ	5
交差検証における Fold 数	5

これらのパラメータをもとに学習したモデルにおいて，

$K = 10, 20, 30, 40, 50$ とし HitRate@K と MRR@K を評価した結果をそれぞれ図 6 と図 7 に示す．これより HitRate@K と MRR@K の 2 つの指標において提案手法がベースラインの精度を上回っていることが分かった．

また，提案手法において各レコードに対する予測スコア上位 50 の企業に正解企業が含まれる時，その正解企業のデータセット内の平均出現回数は 122 回であり，含まれない時，その正解企業のデータセット内の平均出現回数は 13 回であった．このことから出現回数が少ない企業はやはり予測することが難しく HitRate や MRR を下げていることがわかる．

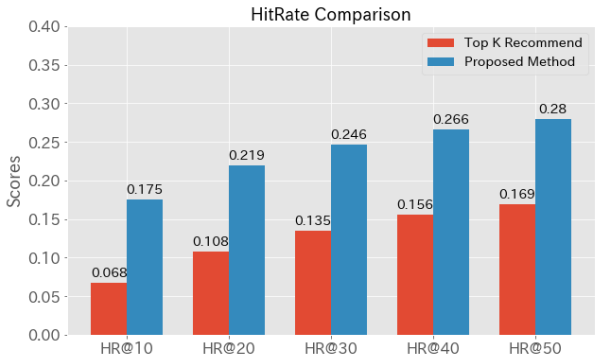

図 6　HitRate@K による評価

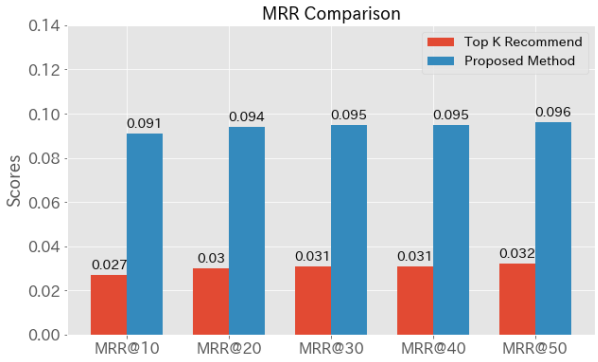

図 7　MRR@K による評価

5　まとめと今後の課題

本研究では，企業と求職者からなるヘテロネットワークを構築し，LSTM モデルを用いて求職者のキャリアパスを考慮したネクスト企業推薦を行う手法を提案している．実験の結果，HitRate@K と MRR@K の 2 つの指標において提案手法がベースラインの精度を大きく上回った．このことから，グラフ表現によって企業間の関係をうまく捉えることができ，さらに系列の学習によって精度良く推薦を行うことが可能であると考える．ただし，提案手法はグラフ全体を構築し Metapath2vec を用いてエンベッディングを行ったベクトルを学習と評価に用いているため，多少リーケージが起きている可能性があると考えており，調査が必要である．

その他に今後の課題としては，以下が挙げられる．

- 系列の学習を用いないモデルと提案手法の精度比較を行う．
- Metapath2vec と LSTM を 2 段階で学習しているが，それらを End-to-end に学習する手法を検討する．
- 各企業のベクトル表現を可視化し，ベクトル表現をうまく獲得できているかどうか確認する．
- データセット内において各企業の出現頻度に大きく差があるため，データバイアスの除去の手法を考案する．
- 本研究ではグラフエンベッディング手法として，ランダムウォークベースの Metapath2vec を用いているが，Graph-NeuralNetwork や GraphSAGE [12] を用いることを検討する．
- より大規模なデータセットを用いて本手法に汎用性があるかどうかを検証する．

これらを検証し、推薦システムの改善に努める．

6　謝　　辞

本研究の一部は科研費（19H04116）と総務省 SCOPE（201607008）による．

文　　献

[1] Linkedin. Linkedin marketing solutions, 2020. https://business.linkedin.com/marketing-solutions.

[2] Guy Berger. Millennials job-hop more than previous generations, they aren’t slowing down, 2016. https://bit.ly/3pgGak9.

[3] Y. Zhang, C. Yang, and Z. Niu. A research of job recommendation system based on collaborative filtering. In *2014 Seventh International Symposium on Computational Intelligence and Design*, Vol. 1, pp. 533–538, 2014.

[4] Ioannis Paparrizos, Berkant Cambazoglu, and Aristides Gionis. Machine learned job recommendation. pp. 325–328, 10 2011.

[5] Snorre S. Frid-Nielsen. Find my next job: Labor market recommendations using administrative big data. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys ’19, p. 408–412, New York, NY, USA, 2019. Association for Computing Machinery.

[6] Le Zhang, Tong Xu, Hengshu Zhu, Chuan Qin, Qingxin Meng, Hui Xiong, and Enhong Chen. Large-scale talent flow embedding for company competitive analysis. In *Proceedings of The Web Conference 2020*, WWW ’20, p. 2354–2364, New York, NY, USA, 2020. Association for Computing Machinery.

[7] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In *KDD ’17*, pp. 135–144. ACM, 2017.

[8] Felix A. Gers, Nicol N. Schraudolph, and Jürgen Schmidhuber. Learning precise timing with lstm recurrent networks. *J. Mach. Learn. Res.*, Vol. 3, No. null, p. 115–143, March 2003.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, p. 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[10] Understanding lstm networks, 2015. http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[11] Linkedin. https://www.linkedin.com/.
[12] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs, 2018.