

## 空間的依存性を考慮したネットワークエンベディング手法の提案

## The Network Embedding Method Considering Spatial Dependency.

大知 正直\*<sup>1</sup> 浅谷 公威\*<sup>1</sup> 森 純一郎\*<sup>1</sup> 坂田 一郎\*<sup>1</sup>  
 Masanao Ochi Kimitake Asatani Junichiro Mori Ichiro Sakata

\*<sup>1</sup>東京大学工学系研究科技術経営戦略学専攻

Graduate School of Technology Management for Innovation, The University of Tokyo

Recently, a few methods to embed the network graph structure into low-dimensional vector spaces are proposed. In this paper, we propose a network embedding method considering spatial dependency, and show the result applied for actual railway passenger movements data in the west part of Japan. We discuss what the role of station is from passenger movements data, and it is important to remove spacial dependency. This proposed method will be useful to infer each passenger movements purpose and discover the undetected role of stations.

## 1. はじめに

2012年、政府は「観光立国推進基本法」を成立し、基本計画を策定した。そこには、訪日外国人旅行者数を2020年までに2,500万人とすることを目標とし、今後日本国内の地域の魅力の発信、インフラ整備に邁進して行くことが掲げられている\*<sup>1</sup>。また、一方で将来の人口減少、高齢化を見据え、公共交通の需要喚起へのさまざまな方策が検討されている。このような状況の中で、地域の魅力、地域の果たす役割をどのように発信するかの一つの課題となっている。この課題を解決するための一つのアプローチとして、データを用いた定量的な分析によるアプローチが考えられる。

これまで、人々の大規模な移動に関する研究は、交通ICカードの登場以来蓄積されたデータを元に様々な研究がなされてきた。代表的なものとしては、障害発生時の需要予測、可視化に関するもの[Yokoyama 14a]や個人の移動目的を推定しようとする試み[Zhang 15]があげられる。また、一方で機械学習の分野では、ネットワーク構造を低次元空間に写像する試みがなされている[Perozzi 14, Tang 15b]。本研究では、移動によって発生する地域間のつながりをネットワーク構造として捉え、ネットワークの中で各地域の果たす役割を明らかにすることを目的とする。

人の移動には様々な地理的な制約が課されている。自身の住んでいる地域から遠く離れた場所へ日常的に移動をしている人は少ないだろう。そして、多くの方は自身の住む地域を中心にさまざまな目的を達成するために、公共交通を利用して近距離移動をしている。だから、交通データは人々が目的を達成するために自分の拠点となる地域を中心に移動した結果の蓄積とみなして良いだろう。では、逆に人々の移動の系列データから買い物するために行く駅、乗り換えのために行く駅というのを検出することは可能だろうか。これはそれほど簡単では無い。言語学における分布仮説では、ある語の出現する周辺の語群が似ているならばその語は近い意味を持つとされている。しかし、人々の移動の系列データの場合、これをそのまま当てはめることは難しい。なぜならば、例えば買い物という目的を達成するために東京の新宿に行く人々の移動データと大阪の梅田に

行く人々とは乗車地点が多くの場合全く異なることが考えられるからである。つまり、新宿と梅田の役割を近いものと考えためには地理的な制約による影響を考慮する必要がある。

そこで、本稿では研究の目的達成の手段として、移動ネットワークから地理的な制約を除去した上で地域の果たす役割を推定する手法を提案する。

本稿では、人々の2地点間の移動は、その人が果たしたい目的と地理的な制約の2つによって発生していると考えられる。そこで、2地点間の移動ネットワーク、地理的近接性を表すネットワーク、少数の地点の果たす役割を示すネットワークを元にすべての地点と役割を低次元空間上に写像する手法を提案する。これまでの異種ノードを写像する手法は、すべてのネットワークにおいて、ノードが写像されるベクトルは同一であったが、本稿で提案する手法では地理的な近接性と役割のネットワークにおいてノードが写像されるベクトルは異なるものとする。

実験は、関西圏の鉄道で日常的に人々が非接触型交通ICカードを用いて移動している履歴を用いて行った。まず、提案アルゴリズムによって写像されたベクトルは、二次元で可視化した図において、地理的な近接性を緩和していることを確認した。また、実験結果では、設定した役割推定のタスクにおいて、提案手法が既存手法を上回る精度を示した。

本稿の貢献を以下にあげる。

1. 地理的な制約が大きな影響を及ぼす場合のネットワークエンベディング手法を提案した。
2. 半教師あり学習の枠組みを利用することで、利用者数の多くない地域の果たす役割についても明らかにすることができる可能性を示した。
3. 提案手法を今後応用できる分野、タスクについて考察した。

2章で関連研究について述べ、本稿の論点を明確にする。次に、本稿で提案する手法について説明を行う。その後、実験に使用するデータについて説明を行う。そして、実験と結果を示し、提案手法が効果を発揮する条件、応用可能性について考察する。最後に結論と今後の展望について述べる。

## 2. 関連研究

ここでは、本研究に関連する研究と本稿で提案する手法に関連する研究、手法について説明する。

連絡先: 大知 正直, 東京大学工学系研究科技術経営戦略学専攻, 〒113-8656 東京都文京区本郷 7-3-1, masanao.oochi@gmail.com

\*<sup>1</sup> <http://www.mlit.go.jp/common/000208713.pdf>

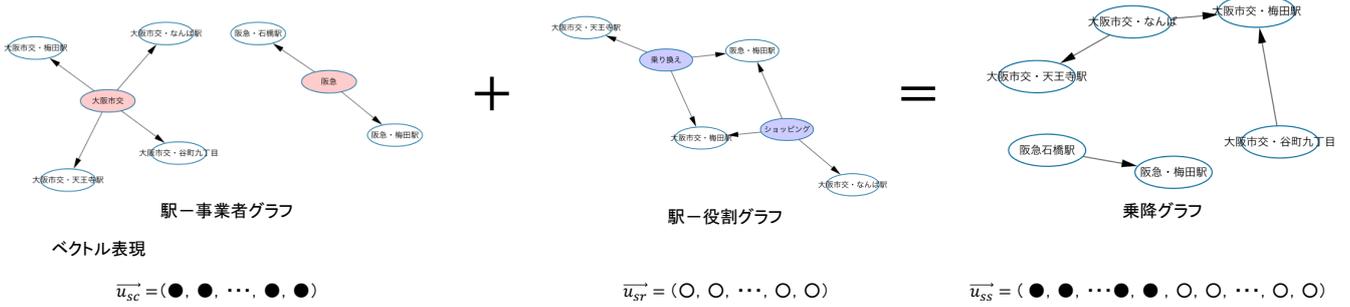


図 1: 提案手法のイメージ図。

## 2.1 人々の大規模な移動に関する研究

人々の大規模な移動に関する研究はすでに様々な形で報告されている。特に、交通 IC カードの普及に伴い、個人の長期間での移動データについて大規模に蓄積されてきており、このデータの分析を行った報告が多くされている。例えば、ロンドンの地下鉄で利用できる、“Oyster” カードの利用データに基づいたものでは、運賃と人々の移動の影響についての分析 [Lathia 11] や郊外と都心における人々の移動構造についての分析 [Roth 11] が報告されている。国内でも “Suica” カードの利用データを利用して、障害発生時の影響を定量化する手法の検証に用いた報告 [角田 13] や “PASMO” のデータを用いた大規模ユーザの移動の可視化 [Itoh 14], 障害発生時の需要予測方法についての報告 [Yokoyama 14b] がある。本稿でも、関西圏の鉄道事業者 6 社局から提供された大規模な人々の移動データを利用する。そして、個別の駅の果たす役割を明らかにするとともに全体として駅の果たす役割、関係性を概観することを目的としている。

## 2.2 ネットワークエンベディングに関する研究

自然言語処理分野から提案された文章の文脈情報を用いて、語を低次元空間へ写像する手法 [Mikolov 13] はその公開されたソフトウェアの手軽さ、速度もあいまって盛んに研究された。そして、最近では、大規模ネットワークに用いることでネットワーク構造を低次元空間へと写像する試みがなされるようになった。Perozzi らは、大規模ネットワーク上でノード間のエッジの重みに応じてランダムウォークさせ、ノードを低次元空間に短時間に写像する手法を提案した [Perozzi 14]。また、ノード間の直接のエッジだけでなく、ノード同士が共通に接続するいわゆる “二次のつながり” のあるノード群も考慮した手法も提案されている [Tang 15b]。これらの手法は教師なし学習の枠組みの中で写像する手法で、同質のノードのみのネットワークだけでなく、一部にラベルや異質なノードも含む、複数のネットワークを元にすべてのノードを同一の低次元空間に写像する、PTE という手法も提案されている [Tang 15a]。

本稿では、人々の移動を地点間のネットワークと捉え、その地点の果たしている役割を推定することを目指している。それには地理的な制約が強く除去する必要がある。そのため、移動ネットワークは地理的なネットワークと地点の役割のネットワークの 2 つのネットワークを合成したものとモデルを提案する。そして提案したモデルにおいてそれぞれのネットワーク上のノードは別々の低次元空間へ写像するように表現学習を行う。

## 3. 提案手法

ここでは提案手法の概要とその詳細について説明する。

### 3.1 概要

本稿では、図 1 に示すようなネットワークモデルを提案する。このモデルは人々の移動ネットワーク (乗降グラフ) は、駅の近接性 (駅-事業者グラフ) と駅の果たす役割 (駅-役割グラフ) の和によって表されると仮定するものである。ここで、駅-事業者グラフは駅の地理的な近接性を表しているもので、同じ路線、事業者に所属する駅であれば地理的に近いということを表したものである。このモデルはつまり、あるユーザが目的を果たすために、現在いる地点の近くから目的を達成できる駅を目指すということを表している。この提案したモデルにおいて、乗降グラフ ( $G_{ss}$ )、駅-事業者 ( $G_{sc}$ ) グラフ、駅-役割 ( $G_{sr}$ ) グラフのそれぞれのノードの持つベクトルはすべて異なる。具体的には、ある駅  $v_i$  の獲得するベクトルは、乗降グラフでは  $\vec{u}_{ss_i}$ 、駅-事業者グラフでは  $\vec{u}_{sc_i}$ 、駅-役割グラフでは  $\vec{u}_{sr_i}$  とする。そしてこれらの間には先に示したモデルによって、以下の式が成り立つものとする。

$$\vec{u}_{ss_i} = \vec{u}_{sc_i} + \vec{u}_{sr_i} \quad (1)$$

ただし、ここでの “+” は加算ではなく 2 つのベクトルを接続して、2 倍の次元数の新たなベクトルとする、という意味である。そして、目的関数  $O$  を式 2 のように設定し、個別の目的関数に対し最適化を行う。

$$O_{all} = O_{sc} + O_{sr} + O_{ss} \quad (2)$$

各グラフに個別に設定する目的関数は以下の式 3 で表す。この式において、 $\hat{p}(\cdot|v_i), p(\cdot|v_i)$  はある地点 (駅)  $v_i$  から別の駅群への移動確率の分布を表す。そして、 $\hat{p}$  は実際の観測結果、 $p$  は写像したベクトル同士の内積をシグモイド関数で確率値に変換したものとする。さらに式中において関数  $d$  は 2 つの分布の距離を表す関数で、今回は KL 情報量を採用する。また、 $\lambda_i$  には観測データを元に算出される、各頂点の次数を採用した。

$$O = \sum_{i \in V} \lambda_i d(\hat{p}(\cdot|v_i), p(\cdot|v_i)) \quad (3)$$

この個別の目的関数を確率的勾配降下法 (SGD) によって最適化を行う。ここで、頂点  $v_i$  を写像するベクトルを頂点ベクトル  $\vec{u}_i$ 、頂点  $v_i$  から遷移する頂点  $v_j$  を写像するベクトルを文脈ベクトル  $\vec{u}_j^i$  とし、頂点ベクトルと文脈ベクトルを次節で提案するアルゴリズムによって学習する。

### 3.2 学習アルゴリズム

定義したベクトル, 更新式を元に表 1 に示すアルゴリズムでベクトル表現の獲得を行う. このアルゴリズムは, 3つのネットワーク  $G_{sc}, G_{sr}, G_{ss}$  それぞれに設定した目的関数の最適化を繰り返し行うものとなっている. また,  $G_{sc}, G_{sr}$  によって更新したベクトルを  $G_{ss}$  のベクトル更新時に合成したものを代入している点が特徴である. このアルゴリズムによって, ノードの一部にしか役割ラベルをタグ付けされていない場合においても, 乗降グラフにおいて関連性の高い駅の場合, 獲得するベクトルの類似度が高いものとなることが期待できる.

表 1: 学習アルゴリズム.

学習アルゴリズム	
1:	各ベクトル表現の初期化
2:	WHILE(設定したサンプリング回数)
3:	$G_{sc}$ からエッジをサンプリング
4:	$u_{ss}, u_{ss}^{\vec{}}$ から $u_{sc}, u_{sc}^{\vec{}}$ 部分を抽出
5:	$u_{sc}, u_{sc}^{\vec{}}$ の更新
6:	$u_{sc}, u_{sc}^{\vec{}}$ から $u_{ss}, u_{ss}^{\vec{}}$ を上書き
7:	$G_{sr}$ からエッジをサンプリング
8:	$u_{ss}, u_{ss}^{\vec{}}$ から $u_{sr}, u_{sr}^{\vec{}}$ 部分を抽出
9:	$u_{sr}, u_{sr}^{\vec{}}$ の更新
10:	$u_{sr}, u_{sr}^{\vec{}}$ から $u_{ss}, u_{ss}^{\vec{}}$ を上書き
11:	$G_{ss}$ からエッジをサンプリング
12:	$u_{ss}, u_{ss}^{\vec{}}$ の更新
13:	END

## 4. 実験に使用するデータ

本稿では実験を 2つのデータを用いて行う. 1つは乗降データで, もう1つは主要駅の利用目的のデータである. ここでは, この2つのデータの説明を行う.

### 4.1 乗降データ

実験には関西圏の鉄道事業者 6 社局から提供された IC カードユーザの乗降データを使用する. このデータは事業者側がデータの匿名化を行っており, 復元できない形のものとなっている. データの内容は, 各ユーザの性別, 年代と乗降日時, 乗降駅となっている. 今回は 2015 年 03 月 01 日~2015 年 03 月 31 日の乗降データを利用した. このデータの概要を表 2 に示す.

表 2: 実験に使用する乗降データの概要.

期間	2015/03/01~2015/03/31
データの総数	50,925,951
ユーザの異なり数	2,007,507
乗降駅の異なり数	672
事業者数	6

### 4.2 主要駅の利用目的データ

本研究では駅の役割の推定を行うことを目的とする. そこで, 利用者数の少ない大多数の駅の利用目的を推定するために, 主要駅に関しては利用者からアンケートを取り, その結果を半教師あり学習の枠組みで用いることで, その他の駅の利用目的を推定するアプローチを取ることとした. 今回, アンケートの取得はクラウドソーシングサービスの一つである Lancers<sup>\*2</sup> を通して行った. アンケート調査は普段から各事業者を利用しているユーザを対象に 2016 年 03 月 03 日から 2 週間行った. アンケートは乗降客数が多い上位 100 駅について, 各事業者ごとに分けて取得している. 質問内容は, 各目的(最

\*2 <http://www.lancers.jp>

寄り, 通勤/通学, 乗り換え, 買い物, 食事, 遊興)を達成するために, その事業者内で行く駅について 3 つずつ答える形で行った. 結果として延べ 223 人のユーザから 96 駅に関する 2,219 のアンケート結果を取得できた. ただし, 実験では(最寄り, 買い物)の 2つのアンケート結果のみを利用している.

## 5. 実験と結果

### 5.1 入力データ

4. 章で述べたデータセットを元に実験で入力するデータの作成を行う. 駅—事業者グラフは, ある駅がどの事業者に所属しているかを表すグラフであり, 無向グラフとし, エッジの重みは 1.0 として入力を行う. 駅—役割グラフは, ある駅の果たしている役割の分布を表すグラフであり, 無向グラフとし, エッジの重みは  $P(\text{役割} | \text{駅})$  とした. つまり, 1つの駅が果たす各役割の割合をすべての役割で合計したものは 1.0 で正規化されている. 最後に乗降グラフは, 駅間の乗降を表したグラフで, 無向グラフとし, エッジの重みは,  $P(\text{降車駅} | \text{乗車駅})$  とした. つまり, 1つの乗車駅から移動し, 降車する駅の割合をすべての降車駅で合計したものは, 1.0 で正規化されている.

### 5.2 実験方法

実験は作成したデータを入力し, PTE 法と提案手法とで学習し, 学習結果を評価することで行う. 評価方法は, t-SNE 法 [Maaten 08] によって学習したベクトルを 2次元空間に次元圧縮を行い可視化したものと, 役割ラベルを推定した結果を元に行う. 推定方法を駅—役割グラフの 2 交差検定によって行う. つまり,  $G_{ss}, G_{sc}$  は訓練時にすべて使用するが,  $G_{sr}$  はランダム抽出した 50% のデータを訓練データに, 残りの 50% はテストデータとし, テストデータに対する推定の結果で比較を行う.

### 5.3 結果

実験の結果, 学習したベクトル表現を 2次元に次元圧縮し, 可視化したものを以下の図 2, 3, 4 に示す. また, 役割ラベルの推定結果を表 3 に示す.

## 6. 考察

図 2,3,4 は, 各社局ごとにそれぞれの駅を色分けし, 同時に獲得した社局, 役割ノードについても同時に可視化したものである. 図 3 を見ると, 図 2 と比較し, 役割ノードが分散して配置されていることが確認される. 観察の結果, 図 3 において, 買い物 (“shopping”) のノードの近くには, 都心の乗降客数の多い駅, 最寄り (“nearest”) のノードには, 郊外の乗降客数の少ない駅が集まっていることが確認された. また, こうした結果は役割の推定精度向上に有効に働いており, 表 3 では PTE 法と比較し, 提案手法 (“proposed(ss)”) が最も高い精度を示していることが確認される.

これらの結果から提案手法は PTE 法と比較して, 空間的依存性の大きい移動データにおいて, 各ノードの果たす役割についてうまく特徴を捉えていることがわかった.

## 7. 結論と今後の展望

本稿では, 駅の役割を表現するベクトルを抽出する手法を提案した. これによって, 駅の役割を推定するタスクにおいて有効に働く可能性があることを示した. しかし, 今回の提案は駅間移動における実験の初期段階の結果を示すのみにとどまっている.

今後は, 提案手法をより洗練させるとともに, 多面的な評価を行う必要があるだろう. また, 手法は複数グラフをまたいだ

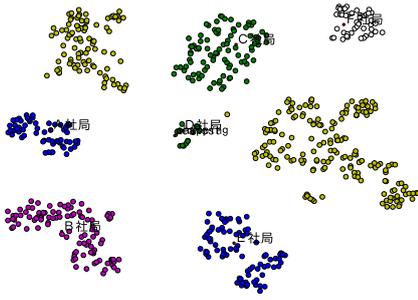


図 2: PTE 法の学習結果.

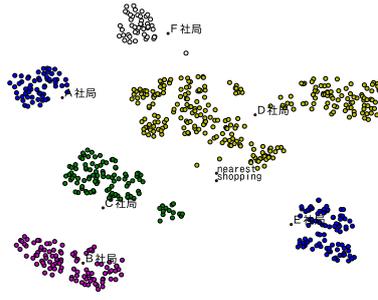


図 3: 提案手法の学習結果 ( $u^{ss}$ ).

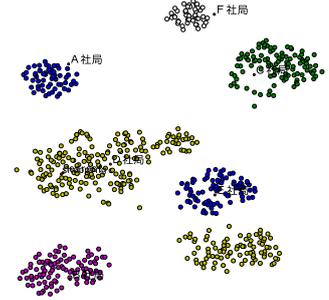


図 4: 提案手法の学習結果 ( $u^{sc}$ ).

表 3: 役割推定の結果 (2 交差検定の平均値).

手法	Macro			Micro		
	Precision(%)	Recall(%)	F1-value(%)	Precision(%)	Recall(%)	F1-value(%)
PTE(joint)	30.556	45.833	36.376	57.190	57.190	57.190
proposed(sc)	31.536	50.000	38.547	63.072	63.072	63.072
proposed(sr)	31.076	47.917	37.500	60.131	60.131	60.131
proposed(ss)	<b>31.536</b>	<b>50.000</b>	<b>38.547</b>	<b>63.072</b>	<b>63.072</b>	<b>63.072</b>

ベクトル表現の同時学習と捉えることができる。その点において、地域や役割以外のさまざまな複数のグラフを盛り込んだモデルや、時系列のグラフ分析へモデルを拡張する可能性が考えられる。

## 参考文献

- [Itoh 14] Itoh, M., Yokoyama, D., Toyoda, M., Tomita, Y., Kawamura, S., and Kitsuregawa, M.: Visual fusion of mega-city big data: An application to traffic and tweets data analysis of Metro passengers, in *Big Data (Big Data), 2014 IEEE International Conference on*, pp. 431–440 (2014)
- [Lathia 11] Lathia, N. and Capra, L.: How Smart is Your Smartcard?: Measuring Travel Behaviours, Perceptions, and Incentives, in *Proceedings of the 13th International Conference on Ubiquitous Computing, UbiComp '11*, pp. 291–300, New York, NY, USA (2011), ACM
- [Maaten 08] Maaten, van der L. and Hinton, G.: Visualizing High-Dimensional Data Using t-SNE (2008)
- [Mikolov 13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, in Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. eds., *Advances in Neural Information Processing Systems 26*, pp. 3111–3119, Curran Associates, Inc. (2013)
- [Perozzi 14] Perozzi, B., Al-Rfou, R., and Skiena, S.: DeepWalk: Online Learning of Social Representations, in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*, pp. 701–710, New York, NY, USA (2014), ACM
- [Roth 11] Roth, C., Kang, S. M., Batty, M., and Barthélemy, M.: Structure of urban movements: polycentric activity and entangled hierarchical flows, *PLoS one*, Vol. 6, No. 1, p. e15923 (2011)
- [Tang 15a] Tang, J., Qu, M., and Mei, Q.: 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*, pp. 1165–1174, New York, NY, USA (2015), ACM
- [Tang 15b] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q.: LINE: Large-scale Information Network Embedding, in *Proceedings of the 24th International Conference on World Wide Web, WWW '15*, pp. 1067–1077, New York, NY, USA (2015), ACM
- [Yokoyama 14a] Yokoyama, D., Itoh, M., Toyoda, M., Tomita, Y., Kawamura, S., and Kitsuregawa, M.: *Advances in Knowledge Discovery and Data Mining: 18th Pacific-Asia Conference, PAKDD 2014, Tainan, Taiwan, May 13-16, 2014. Proceedings, Part I*, chapter A Framework for Large-Scale Train Trip Record Analysis and Its Application to Passengers' Flow Prediction after Train Accidents, pp. 533–544, Springer International Publishing, Cham (2014)
- [Yokoyama 14b] Yokoyama, D., Itoh, M., Toyoda, M., Tomita, Y., Kawamura, S., and Kitsuregawa, M.: A framework for large-scale train trip record analysis and its application to passengers' flow prediction after train accidents, in *Advances in Knowledge Discovery and Data Mining*, pp. 533–544, Springer (2014)
- [Zhang 15] Zhang, F., Yuan, N. J., Wang, Y., and Xie, X.: Reconstructing Individual Mobility from Smart Card Transactions: A Collaborative Space Alignment Approach, *Knowl. Inf. Syst.*, Vol. 44, No. 2, pp. 299–323 (2015)
- [角田 13] 角田 史記, 加藤 学, 大塚 理恵子, 助田 浩子, 大関 一博: 交通系 IC カードを利用した鉄道輸送障害時の影響を定量化する方法の研究, *情報処理学会論文誌データベース (TOD)*, Vol. 6, No. 3, pp. 187–196 (2013)