Vol.2018-MPS-121 No.11
2018/12/18

就職ポータルサイト上の行動履歴データに基づく企業の分散表現モデルに関する一考察

杉山 裕貴[†1,a)]　雲居 玄道[†1]　後藤 正幸[†1]　桜井 崇[†2]

概要：近年，採用活動を行う企業や就職活動を行う学生の多くが就職ポータルサイトを利用しており，サイト運営企業は、サイト上でのユーザの行動履歴を蓄積している．本研究では，ユーザの行動履歴データに自然言語処理モデルの Word2Vec を用いて，就職ポータルサイト上の企業の分散表現を得る方法について検討する．Word2Vec を行動履歴データに適用した場合，複数の行動を組合せた分析によりユーザの嗜好をより的確に捉えられる可能性がある．そこで本稿では，複数のエントリーの組み合わせを考慮した分散表現モデルを提案し，大手就職ポータルサイトの実データに適用することで，その有効性を検証する．

1. はじめに

近年，採用活動を行う企業や就職活動を行う学生（以下，ユーザ）の多くが就職ポータルサイトを利用している．企業は就職ポータルサイト上で自社の基本情報や採用情報を個社ページに掲載し，ユーザからのエントリーを募ることができる．一方，ユーザは掲載企業の個社ページや業界・仕事研究の記事等を閲覧することで企業や業種の魅力を知り，興味のある企業へエントリーをすることができる．就職ポータルサイト運営会社は，サイトを通じて就職活動を行うユーザの膨大な行動履歴データを分析し，掲載企業への施策提案やサイトの改善に活かすことが可能である．

また，池田の報告 [1] では，自然言語処理モデルの 1 つである Word2Vec[2] を，本研究が対象とする就職ポータルサイト等複数の Web サービスにおける推薦に適用することで，コンバージョン率が向上することが示されている．Word2Vec は文章中の単語を低次元空間上の点として表現することを可能にし，これを「単語分散表現モデル」と呼ぶ．この事例では，Word2Vec をユーザの行動履歴に適用し，Web サービス上のアイテムを低次元の空間上の点として表現，それらの類似度を算出し，ユーザが行動をとったアイテムと類似度の高いアイテムを志向に合致するアイテムと考え，推薦候補としている．しかし，ユーザへ同時に提示できるアイテム数には限りがある中で，ユーザが行動をとった複数アイテムの組合せに対して類似度を算出することで，1 対 1 での類似度算出よりも的確にユーザの嗜好を捉えた推薦候補の決定ができる可能性がある．

そこで本稿では，大手就職ポータルサイトにおける複数企業の組合せを 1 つの要素として扱う分散表現モデルを提案する．また，提案手法を実データに適用し，1 社単位と 2 社の組合せでの企業間類似度算出の結果の比較により，その有効性を検証する．

2. アイテムの分散表現に基づく推薦システム

Web サービスにおける推薦システムでは従来，行列分解を用いた手法 [3] などが用いられてきた．近年では，購買履歴や Web サービスのデータに自然言語処理の手法である Word2Vec を適用した事例 [1],[4] などが報告され，推薦における有用性が示されている．これらの事例では，各ユーザの行動履歴を 1 文章，行動対象のアイテムを単語と置き換えて，Word2Vec を適用することで各アイテムの低次元の分散表現を獲得する．そして，得られた分散表現からアイテム間の類似度を算出し，類似度の高いアイテムをユーザの興味に合致するものとして推薦する．

しかし，ユーザが行動した複数アイテムの組合せに着目した分散表現の学習及び類似度算出を行うと，1 対 1 のアイテム間類似度算出結果とは異なる傾向を示す可能性がある．就職ポータルサイトのエントリー履歴の例を考えると，業種の異なる A 社と B 社に対してはそれぞれ同業種の企業が高い類似度を示しやすいが，2 社両方をエントリーしたユーザにとってはグループ企業や企業風土といった，業種以外の嗜好の軸が存在するという可能性がある．この場合，ユーザのエントリー企業の組合せに着目した分散表

[†1] 現在，早稲田大学
Presently with Waseda University, Shinjuku, Tokyo 169–0072, Japan
[†2] 現在，株式会社 リクルートキャリア
Presently with Recruit Career Co., Ltd.
[a)] miyus0919yuuki@ruri.waseda.jp

ⓒ 2018 Information Processing Society of Japan

現の学習を行い，A 社と B 社の組合せに対して類似度の高い企業の組合せの中から推薦候補を決定することで，よりユーザの嗜好を捉えた推薦が実現できる可能性がある．

3. 提案手法

同一ユーザによってエントリーされた複数の企業の共起性に着目し，企業の組合せ間での類似度算出を行うために，各ユーザのエントリー履歴から 2 社の組合せをそれぞれ生成し，2 社の組合せからなるユーザごとのエントリー系列に対して Word2Vec を適用する．

Word2Vec への入力データは，各ユーザごとに N 件のエントリー履歴から $\binom{N}{2}$ 通りの 2 社の組合せを生成し，ランダムに並び替えたものを各ユーザの組合せ系列として使用する．このとき，2 社の組合せを 1 単語，各ユーザの $\binom{N}{2}$ 通りの組合せをランダム置換した系列を 1 文章と置き換えて Word2Vec を適用し，2 社の組合せの分散表現を獲得する．これにより，2 社の組合せ同士での類似度算出が可能になる．

4. 分析

4.1 分析概要

複数の企業の組合せを 1 つの要素として扱う分散表現モデルの適用で，1 社単位での企業間類似度算出と異なる結果を示し，よりユーザに合った推薦候補企業決定への適用可能性を示すため，大手就職ポータルサイト A の実データを用いた分析を行う．分析対象データは，2015 年 3 月 31 日 23 時 59 分 59 秒の時点での各ユーザの直近 10 件のエントリー履歴とする．これは，推薦企業リストを作成するタイミングに近い時期のエントリー履歴を用いることで，時期ごとに変化しうるユーザの行動傾向を適切に捉えるためである．また，本分析では総エントリー数が 10 件に満たないユーザのエントリー履歴データは分析対象としない．

また事前分析の結果，1 社単位の Word2Vec のベクトルの次元数を 20，ウィンドウサイズを 3，2 社の組み合わせの Word2Vec のベクトルの次元数を 25，ウィンドウサイズを 5 とした．また，両モデルで共通してネガティブサンプル数を 10 とし，skip-gram モデルを用いた．

4.2 分析結果と考察

1 社単位の Word2Vec と提案手法で得られた企業の分散表現の類似度算出結果を比較する．本稿では，C 社，D 社（業種：自動車）と E 社（業種：輸送機器）の分析例を示す．**表 1** から**表 3** は C 社，D 社，C 社&D 社，C 社&E 社と類似度の高い企業及び企業の組合せの上位 5 件である．

C，D，E 社単体については，表 1 のように，それぞれ同業種の企業を中心に高い類似度を示している．C 社と D 社の組合せに着目すると，表 1 のように，自動車の企業を中心に高い類似度を示し，このエントリーの組合せからは

表 1 C 社，D 社，E 社と cos 類似度の高い企業上位 5 件

C 社と高類似度	D 社と高類似度	E 社と高類似度
D 社（自動車）	H 社（自動車）	M 社（自動車）
E 社（輸送機器）	F 社（自動車）	N 社（自動車）
J 社（輸送機器）	G 社（自動車）	O 社（輸送機器）
G 社（自動車）	I 社（自動車）	J 社（輸送機器）
H 社（自動車）	C 社（自動車）	N 社（輸送機器）

表 2 C 社&D 社と cos 類似度の高い企業の組合せ上位 5 件

企業（業種）		cos 類似度
C 社（自動車）	H 社（自動車）	0.922
E 社（自動車）	C 社（自動車）	0.913
G 社（自動車）	C 社（自動車）	0.912
K 社（自動車）	C 社（自動車）	0.912
L 社（総合電機）	C 社（自動車）	0.906

表 3 C 社&E 社と cos 類似度の高い企業の組合せ上位 5 件

企業（業種）		cos 類似度
C 社（自動車）	M 社（自動車）	0.890
J 社（輸送機器）	C 社（自動車）	0.846
E 社（輸送機器）	M 社（自動車）	0.843
E 社（輸送機器）	D 社（自動車）	0.824
N 社（輸送機器）	C 社（自動車）	0.809

表 2 より「自動車」という志向が読み取れる一方，総合電機の L 社という C，D 社単体の類似度算出では上位に出現しない企業が現れた．また，C 社と E 社の組合せに着目すると，表 3 のように自動車や輸送機器の企業が高い類似度を示しており，実際の企業概要を確認すると，「C 社中心の企業グループに属する企業」という嗜好が読み取れる．つまり，C 社にエントリーしたユーザの中でも，他に D 社か E 社エントリーしたかどうかで，異なる嗜好があることが推定され，推薦すべき企業が変わってくると考えられる．以上のことより，2 社の組合せに着目した企業の分散表現から，ユーザの嗜好をより細かく捉えることが可能になる．

5. まとめと今後の課題

本稿では，ユーザにエントリーされた企業の共起性に着目した分散表現モデルを提案し，企業間類似度算出の結果の比較より，よりユーザに合った推薦企業決定への適用可能性を示した．また，今後の課題として，エントリーの組合せを考慮した推薦システムの検討などが挙げられる．

参考文献

[1] 池田 裕一：リクルート式 自然言語処理技術の適用事例紹介，*WebDB Forum 2016*(2016)
[2] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient Estimation of Word Representations in Vector Space. *arXiv preprint* arXiv:1301.3781(2013).
[3] Koren, Y., Bell, Y., Volinsky: Matrix Factorization Techniques for Recommender Systems. *IEEE Computer*, 42(8), 30–37(2009)
[4] Phi, V.T.,Liu, C. and Hirate, Y.: Distributed Representation-based Recommender Systems in E-commerce. *DEIM Forum 2016*, C8-1(2016).

ⓒ 2018 Information Processing Society of Japan