就職ポータルサイトにおけるユーザのエントリー履歴に基づく企業の分散表現モデルに関する研究

情報数理応用研究　5217C021-9　杉山裕貴
指導教員　後藤正幸

A Study on a Distributed Representation Model of Companies Based on Users' Entry Histories on a Portal Site for Job-hunting

SUGIYAMA Yuuki

1　研究背景・目的

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．企業は採用広報活動の一環として，就職ポータルサイト上に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集することができる．一方，ユーザは掲載企業の個社ページや業界・仕事研究の記事などの就職活動関連情報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エントリー履歴や個社ページ閲覧履歴など，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．例えば，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の Web サービスにおける推薦に適用することで，エントリーや購買といったコンバージョン率が向上することが示されている．Word2Vec は文書中の単語を低次元空間上の点として表現する言語モデルであり，この表現を「単語分散表現」と呼ぶ．この事例では，Word2Vec をユーザの現時点での嗜好が現れるように直近数件の行動履歴に適用し，就職ポータルサイト上の企業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，ユーザが直近に行動をとった企業と類似度が高いものを嗜好に合致する企業として，推薦候補としている．しかし，ユーザへ同時に提示できる企業数には限りがある中で，ユーザが近い時期に行動をとった複数企業の組合せに着目することにより，1 社単位の類似度では測ることが困難なユーザの嗜好性を評価した推薦ができる可能性がある．すなわち，複数企業の組合せについて分散表現の学習を行い，企業の組合せ間の類似度を算出することで，大企業などは様々な嗜好の軸でエントリーされやすいため，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期待できる．

そこで本研究では，大手就職ポータルサイトにおける複数企業のエントリーの組合せを 1 つの要素として扱うことで，ユーザの行動の共起性に着目した分散表現モデルを提案し，多様な嗜好の軸を反映した企業の分散表現の獲得を可能にする．また，提案手法を実データに適用し，1 社単位で学習した企業の分散表現と 2 社の組合せで学習した企業ペアの分散表現の企業間類似度算出の結果から，ある企業へのエントリーの背景にある嗜好の軸を分析する．池田の従来手法 [1] と提案手法の分析結果の比較を通して，提案手法の有効性を検証する．

2　準備

2.1　就職ポータルサイトを用いた就職活動

近年，日本における新卒学生の就職活動において，Web サービスの利用が盛んになっている．企業は Web サイト上で自社の基本情報・新卒採用情報の掲載や説明会などのイベントへの参加募集，採用選考へのエントリーの受付などを行っている．それに対し，学生は Web サイトを通じて情報収集やイベントおよび採用選考へのエントリーを行う．このような，企業の採用活動と，学生の就職活動を Web 上でサポートするために運営されているのが就職ポータルサイトである．就職ポータルサイトの利用により，企業はユーザに自社への認知・関心を持ってもらう機会が増加し，学生ユーザは自分の興味に合致する企業の発見やエントリーを効率的に進めることが可能になるため，多くの企業と学生ユーザに利用されている．

本研究では，大手就職ポータルサイト A（以下，サイト A）におけるユーザの行動履歴データを対象事例として扱う．ユーザがサイト A で企業へのエントリーなどを行うためにはアカウント登録が必要であり，各ユーザの行動履歴データはユーザごとに独自の ID に紐付いている．そのため，個別のユーザ単位での行動分析や企業の推薦が可能になっている．

2.2　Word2Vec[2]

単語分散表現を学習する手法の代表的なものとして Word2Vec がある．Word2Vec では，「ある単語の意味はその単語の周辺に現れる単語 (文脈) によって与えられる」という仮説のもと，文の集合を入力とし，文中の注目単語を周辺の単語から予測するニューラルネットワークを学習する．そして，このニューラルネットワークの中間層を単語分散表現として出力する．得られた分散表現空間上の位置関係によって，単語間の意味的な類似性を Cos 類似度などの尺度を用いて定量的に算出することが可能となる．また，得られた単語の分散表現を用いて，“queen”-“woman”+“man”=“king”のような単語間での加算，減算による意味の表現も可能である．そのため，類義語の抽出や文書分類のタスクなどに活用されている．

2.3　就職ポータルサイトのデータを用いた関連研究

これまで，就職ポータルサイト上でのユーザの行動履歴や企業のデータに統計学的手法や機械学習手法を用いた研究が行われている．例えば，掲載企業の被エントリー数の予測及び影響要因分析のためのモデルに関する研究

[4] や，企業のアピールポイントとユーザの志望理由の関係性に着目したマッチング分析モデルの構築 [5] などが行われ，実際のサービスへの活用可能性が示されている．

2.4 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，潜在意味解析や非負値行列分解に代表される行列分解を用いた手法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] などが報告され，推薦における有用性や行列分解に比べた計算コストの軽減が示されている．これらの事例では，各ユーザの行動履歴を 1 文書，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嗜好に合致するものとして推薦する．

3 提案手法

3.1 概要

現在，サイト A では Word2Vec をユーザのエントリー履歴に適用してサイト上に掲載されている企業の分散表現を獲得し，ユーザが直近にエントリーした企業と類似度が高い企業をそのユーザの嗜好に合致する企業として推薦を行っている．その際，分散表現の学習は 1 社を 1 単語と置き換えた Word2Vec で行い，企業間の類似度も 1 対 1 で算出している．一般に，1 社単位の分散表現では，同業種の企業が空間上で近い位置に配置され，これらの類似性が高くなる傾向がある．すなわち，業種の異なる A 社と B 社を共にエントリーしたユーザに対しては，A 社の業種の企業と B 社の業種の企業の類似度が高くなり，これらが推薦され易くなる．しかし，業種の異なる A 社と B 社単体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 社両方を近い時期に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可能性がある．例えば，あるユーザが，生命保険 A とその子会社である情報システム B の 2 社をエントリーしていた場合，このユーザは，生命保険 A のグループ企業を志望していることが考えられる．一方，他のユーザが，生命保険 A の子会社である情報システム B と建設 C の子会社である情報システム D の 2 社にエントリーしていた場合は，このユーザは，親会社の業種に関わらず，情報システム関連企業を志望していることが伺える。そこで，ユーザが近い時期に行動した複数企業の共起性に着目した分散表現を獲得できれば，1 対 1 の企業間類似度算出結果とは高い類似度を示す企業の傾向が異なる可能性がある．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現の学習を行い，2 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企業の推薦が実現できる可能性がある．そこで本研究では，同一ユーザによってエントリーされた複数の企業の組合せの分散表現を学習する手法を提案する．

3.2 ユーザがエントリーした企業の共起性に着目した分散表現モデル

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．3 社以上の組合せを生成した場合，企業の共起パターンが急増し，限られたエントリー履歴データからでは分散表現の十分な学習が困難になることから，本研究では 2 社の組合せに限定する．

Word2Vec への入力データは，各ユーザごとに直近 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，組合せの順序による学習の偏りを避けるためランダムに並び替えた系列を 1 文書と置き換えて Word2Vec を適用し，2 社の組合せの分散表現を獲得する．この手法で対象となるエントリー履歴の範囲を限定し，$\binom{N}{2}$ 通りの 2 社の組合せの分散表現を獲得することで，連続したエントリーに限らず，比較的近いタイミングでのエントリーの共起性に基づく企業間の関係性を表現でき，2 社の組合せ同士での類似度算出が可能になる．

4 実データ分析

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 2016 年卒学生向けサイトの各ユーザ直近 10 件のエントリー履歴とした．これは，推薦企業リストの作成を想定した際，推薦を行う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，時期ごとに変化し得るユーザ全体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．

事前分析の結果，Word2Vec の代表的なパラメータを以下のように設定した．また，いずれの手法においても，skip-gram モデル [2] を用いた．

表 1.Word2Vec の主なパラメータ設定

入力データ	1 社単体	2 社組合せ
ベクトルの次元数	20	25
ウィンドウサイズ	3	45
ネガティブサンプル数	10	10
最小出現数の閾値	5	5
エポック数	10	10

2 社の組合せを入力とする提案手法ではユーザのエントリー系列内の順序には意味を持たせず，エントリーした企業およびそれらの組合せの集合として捉えるため，ウィンドウサイズは各手法のエントリー系列の長さと等しく設定した [3]．また，最小出現数の閾値は，極端に頻度の少ないエントリー企業の組合せの分散表現の学習を行わず，学習コストを削減するために設定した．

4.2 1 社単体の Word2Vec と提案手法の比較

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．類似度は Cos 類似度を用いる．本稿では，いずれも業界最大手の自動車 A，

自動車 B と輸送機器 A の 3 社に関する分析例を示す．また，本節では考察のため，同じ都道府県に所在する，自動車 A を中心とした企業グループに所属する企業を下線で強調して表記する．表 2 から表 4 はそれぞれ，従来の 1 社単体の Word2Vec で得られた自動車 A，自動車 B，輸送機器 A と Cos 類似度の高い企業の上位 10 件である．

表 2. 自動車 A と Cos 類似度の高い企業上位 10 件

順位	企業	Cos 類似度
1	自動車 B	0.944
2	輸送機器 A	0.942
3	輸送機器 B	0.937
4	自動車 D	0.903
5	自動車 E	0.898
6	自動車 C	0.897
7	自動車 H	0.895
8	自動車 I	0.893
9	自動車 J	0.877
10	輸送機器 C	0.871

表 3. 自動車 B と Cos 類似度の高い企業上位 10 件

順位	企業	Cos 類似度
1	自動車 E	0.980
2	自動車 C	0.974
3	自動車 D	0.947
4	自動車 F	0.944
5	自動車 A	0.944
6	自動車 K	0.933
7	自動車 G	0.930
8	自動車 L	0.927
9	輸送機器 A	0.926
10	輸送機器 D	0.918

表 4. 輸送機器 A と Cos 類似度の高い企業上位 10 件

順位	企業	Cos 類似度
1	自動車 H	0.981
2	自動車 J	0.968
3	輸送機器 C	0.957
4	輸送機器 B	0.956
5	自動車 M	0.943
6	自動車 A	0.942
7	輸送機器 E	0.937
8	自動車 B	0.926
9	自動車 N	0.918
10	自動車 E	0.907

表 2 から表 4 より，自動車 A，自動車 B，輸送機器 A 単体については，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．その中でも，自動車 A と類似度の高い企業には，自動車 A を中心とした企業グループに属する企業と，それ以外の業種が「自動車」の企業が混在している一方，輸送機器 A と類似度の高い企業の多くは，自動車 A を中心とした企業グループに属する企業であった．

表 5，表 6 は（自動車 A, 自動車 B），(自動車 A, 輸送機器 A）の組合せと Cos 類似度の高い企業の組合せ上位 20 件である．2 社の組合せについては，(A 社, B 社）の組合せに対して（A 社, C 社），(B 社, C 社）という組合せが高い類似度を示すように，類似度上位の企業が重複することが考えられる．そのため，1 社単体の企業間類似度算出結果よりも多い，上位 20 件を取得した．

まず，(自動車 A, 自動車 B）の組合せに着目すると，表 5 のように，自動車の企業を中心に特に高い類似度を示し，次いで，自動車に限らない様々な分野のメーカーが高い類似度を示す．また，実際の企業概要より，これらのメーカーは各分野内で比較的規模の大きい企業であった．このことより，自動車 A と自動車 B へのエントリーの組合せからは，「自動車」という業種への嗜好や，大手メーカーへの嗜好が読み取れる．よって，自動車 A と自動車 B を共にエントリーしたユーザには，自動車 A 単体に対して高い類似度を示した輸送機器 A, B よりも，自動車 A と自動車 B の組合せに対して高い類似度を示した自動車 C, D, E, G などを優先的に推薦すべきだと考えられる．

表 5.（自動車 A, 自動車 B）の組合せと Cos 類似度の高い企業の組合せ上位 20 件

順位	企業		Cos 類似度
1	自動車 C	自動車 A	0.967
2	自動車 G	自動車 A	0.965
3	自動車 E	自動車 A	0.965
4	自動車 D	自動車 A	0.951
5	自動車 D	自動車 B	0.947
6	自動車 C	自動車 B	0.934
7	自動車 F	自動車 A	0.933
8	自動車 G	自動車 B	0.931
9	自動車 E	自動車 B	0.921
10	輸送機器 A	自動車 B	0.916
11	重電・産業用電気機器 A	自動車 A	0.910
12	自動車 I	自動車 A	0.909
13	総合商社 A	自動車 B	0.908
14	機械 A	自動車 B	0.905
15	医療機器 A	自動車 B	0.902
16	食品 A	自動車 B	0.901
17	重電・産業用電気機器 B	自動車 B	0.899
18	自動車 C	自動車 E	0.899
19	総合電気 A	自動車 B	0.899
20	自動車 I	自動車 B	0.898

表 6.（自動車 A, 輸送機器 A）の組合せと Cos 類似度の高い企業の組合せ上位 20 件

順位	企業		Cos 類似度
1	自動車 H	自動車 A	0.967
2	自動車 B	輸送機器 A	0.953
3	自動車 H	輸送機器 A	0.927
4	自動車 J	自動車 A	0.922
5	輸送機器 B	自動車 A	0.919
6	輸送機器 B	輸送機器 A	0.915
7	自動車 M	自動車 A	0.915
8	精密機器 A	自動車 A	0.909
9	自動車 D	輸送機器 A	0.906
10	自動車 E	輸送機器 A	0.903
11	重電・産業用電気機器 A	輸送機器 A	0.902
12	輸送機器 C	自動車 A	0.897
13	自動車 M	輸送機器 A	0.895
14	総合電機 B	輸送機器 A	0.894
15	自動車 H	自動車 B	0.891
16	輸送機器 B	自動車 B	0.891
17	機械 A	自動車 A	0.890
18	自動車 J	輸送機器 A	0.890
19	石油・石炭 A	輸送機器 A	0.889
20	自動車 D	自動車 A	0.888

一方で，自動車 A と輸送機器 A の組合せに着目すると，表 6 のように，自動車 A を中心とした企業グループに属する企業を中心に高い類似度を示している．すなわち，自動車 A と輸送機器 A へのエントリーの組合せから

は,「自動車 A を中心とした企業グループ」という嗜好が読み取れる．よって，自動車 A と輸送機器 A を共にエントリーしたユーザには，自動車 A 単体に対して高い類似度を示した自動車 C, D, E などよりも，自動車 A と輸送機器 A の組合せに対して高い類似度を示した自動車 H, G など，自動車 A を中心とした企業グループの企業を優先的に推薦すべきだと考えられる．

以上より，自動車 A にエントリーしたユーザの中でも，その他に自動車 B か輸送機器 A に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．1 対 1 での類似度算出では，自動車 A と類似度の高い企業からは混在した嗜好しか読み取ることができなかったが，2 社の組合せに着目した企業の分散表現の学習と類似度算出によって，ユーザの嗜好の軸をより細かく捉えることが可能になった．

4.3 1 社単体の分散表現の加算と提案手法の比較

次に，1 社単位で学習した企業の分散表現を 2 社分足し合わせたものと，2 社の組合せで学習した企業組合せの分散表現の類似度算出結果の比較を行う．表 7，表 8 はそれぞれ，1 社単位の Word2Vec によって得られた自動車 A と自動車 B，自動車 A と輸送機器 A の分散表現を足し合わせたものと類似度の高い企業上位 15 件である．

表 7. 自動車 A と自動車 B を足し合わせた分散表現と Cos 類似度の高い企業上位 15 件

順位	企業	Cos 類似度
1	自動車 E	0.952
2	自動車 C	0.949
3	輸送機器 A	0.947
4	自動車 D	0.938
5	自動車 F	0.917
6	自動車 K	0.914
7	自動車 I	0.913
8	輸送機器 B	0.905
9	自動車 L	0.900
10	機械 A	0.898
11	自動車 G	0.898
12	自動車 H	0.894
13	輸送機器 D	0.894
14	総合電機 C	0.884
15	輸送機器 F	0.884

表 5 と表 7 を比較すると，2 社の組合せを学習した場合は自動車 A と自動車 B へのエントリーに対して「自動車」という業種の企業が特に高い類似度を示した．一方で，2 社の分散表現を足し合わせた場合は,「自動車」の企業に加え輸送機器の企業および自動車 A を中心とした企業グループに属する企業が複数高い類似度を示している．すなわち，2 社の足し合わせでは自動車 A を中心とした企業グループに属する企業と業種が「自動車」の企業という 2 つの異なる嗜好によってエントリーされる企業が混在したままであった．表 6 と表 8 の比較においても，表 5 と表 7 の場合と同様に 2 つの異なる嗜好によってエントリーされる企業が混在したままであった．

以上より，提案手法によって，1 社単位の分散表現モデルでは捉えにくい，ユーザがエントリーした企業の共起性に着目した分散表現の学習が行えることが明らかとなった．

従来手法と比較した分析により，提案手法は，1 社単位の類似度算出では嗜好の理由が混在しやすい企業，例えば，ユーザにより全く異なる嗜好の軸でエントリーされている大企業との類似度算出において特に有効な手法であると考えられる．

表 8. 自動車 A と輸送機器 A を足し合わせた分散表現と Cos 類似度の高い企業上位 15 件

順位	企業	Cos 類似度
1	輸送機器 B	0.960
2	自動車 H	0.952
3	自動車 B	0.949
4	自動車 J	0.936
5	輸送機器 C	0.928
6	自動車 M	0.917
7	自動車 E	0.916
8	自動車 C	0.909
9	輸送機器 E	0.905
10	自動車 N	0.894
11	機械 A	0.894
12	自動車 D	0.893
13	自動車 F	0.888
14	自動車 I	0.888
15	コンピュータ・通信機器・ＯＡ機器 A	0.888

5 考察

提案手法を実データに適用することで，同じ企業でも共にエントリーされる企業によって類似度が高い企業の傾向が異なる場合があり，その際ユーザが企業にエントリーする際の嗜好の軸を考慮することの重要性が示唆された．

提案手法は，ユーザのエントリー間隔が比較的短い就職活動初期には有用であるが，エントリー間隔が比較的長くなる中期以降は，近いタイミングでのエントリーの共起性がデータに現れにくいことが考えられる．従来手法を適用すべきケースとの使い分けについてはさらに検討の余地がある．

6 まとめと今後の課題

本研究では，ユーザにエントリーされた複数企業の共起性に着目した分散表現モデルを提案し，従来の企業の分散表現の獲得手法との企業間類似度算出の結果の比較により，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重視した推薦システムの構築や，閲覧履歴や個社ページの企業紹介文に Word2Vec を適用することで得られる企業の分散表現の分析などが挙げられる．

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] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” *arXiv preprint*, arXiv:1301.3781, 2013.

[3] 名渡山 夏子, 岡本 一志,“Word2Vec に基づく購買履歴からのアイテムベクトル学習,” 知能と情報, Vol.29, No.3, pp.579–585, 2017.

[4] 永森 誠矢, 山下 遥, 荻原 大陸, 後藤 正幸,“混合回帰に基づく就職ポータルサイトの被エントリ数分析モデルに関する一考察,” 情報処理学会誌, Vol.59, No.4, pp.1273–1285, 2018.

[5] 坂元 哲平, 山下 遥, 荻原 大陸, 後藤 正幸,“就職ポータルサイトにおける企業のアピールポイントと志望理由のマッチング分析モデルに関する一考察,” 情報処理学会誌, Vol.58, No.9, pp.1535–1548, 2017.