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学生の志望理由の関係分析モデルの研究

1X12C051-0　坂元 哲平
指導教員　後藤 正幸

1　研究背景・目的

近年，企業の採用活動や学生の就職活動において，就職ポータルサイトが活用されている．就職ポータルサイトを通して，企業は自社の広報活動を行い，学生からのエントリーを募ることができ，学生は多種多様な企業を志望することが可能である (本研究では「エントリー」だけでなく,「企業に興味を持つこと」等を含めて広く「企業を志望する」事象として扱う)．企業は広報活動の一環として，就職ポータルサイト内の個社ページで自社の基本情報，採用情報，アピールポイント等を掲載することで，就職活動中の学生に対して広く企業情報を提供する．また，学生は得られた企業情報をもとに，企業に魅力を感じ，それが理由となってその企業へと志望する場合があると考えられる．このように，採用活動・就職活動は企業と学生のある種のマッチング活動と捉えることができ，就職ポータルサイトがそのマッチングをサポートする機能を提供していると考えられる．

一方，企業が自社の強みとして掲載するアピールポイントと，学生が感じるその企業の魅力 (企業に対する志望理由) は必ずしもマッチしていない可能性がある．例えば，企業が技術力をアピールポイントとしているにも関わらず，学生は給与水準を志望理由としている場合や，そもそもアピールポイントが学生の志望に影響を与えていない場合である．このようなミスマッチは，選考段階や採用後の企業と学生の相性不良に繋がるとも考えられるため，何らかの解決策が望まれている．そのため，企業のアピールポイントに対して学生がどう魅力を感じ，企業を志望するかを分析し，具体的な問題点を明らかにする必要がある．

そこで本研究では，企業のアピールポイントと学生の志望理由の関係性を分析するためのモデルを提案する．その際，両者の関係性の背景には，各企業特有の情報やアピールの仕方，また，学生ごとに異なる嗜好が混在していると考えられるため，潜在クラスモデルを導入する．さらに，提案モデルを用いて，大手就職ポータルサイト (以下，就活サイト A) における実データを分析し，提案モデルの有効性を示す．

2　準備

2.1　対象事例

本研究で対象とする就活サイト A では，企業は就活サイト A の運営会社が定めたアピールポイント (以下，アピール) 群の中から規定数までアピールを選択し，それに沿った企業紹介文や画像を個社ページ内で掲載することができる．また，当該サイトには，学生が企業を志望する際，その志望理由 (以下，理由) を定められた理由群の中から複数選択可で登録 (以下，理由登録) する機能がある．すなわち，企業のアピールと学生の理由に関するデータが蓄積され，そのデータを分析に活用することができる．

2.2　潜在クラスモデル

潜在クラスモデルは，観測されたデータの背後に観測できない潜在的な変数の存在を仮定したモデルである．潜在的な変数の仮定により，異質のデータが混ざったような現実問題の分析が可能となり，文書分類や購買履歴分析等に適用されている．潜在クラスモデルには，データがある潜在クラスのもとで生起すると仮定する Unigram Mixture(以下，UM)[1] や，データが複数の潜在クラスから確率的に生起することを許容する PLSA[2] がある．本研究対象では，学生が企業にどう魅力を感じるかは理由登録ごとに異なり，また，一回の理由登録においては，アピールと理由がある関係性をもとに同時生起していると考えられる．これを表現するため，PLSA と UM の双方の特徴を有したモデルを提案する．

3　提案モデル

3.1　概要

本研究では，企業のアピールと学生の理由の関係性を分析するために，各理由登録ごとに潜在クラスを仮定する．そこで，一回の理由登録に付随するアピールと理由は，ある潜在クラスのもとで同時に生起するという仮定を置く．一方，学生は複数の企業を志望するため，そのたびに潜在クラスが異なることを許容したモデルの構築が必要となる．そこで，各理由登録に対してアピールと理由がそれぞれベクトルで与えられることに対応した潜在クラスモデルを新たに提案する．さらに，パラメータを推定する際の確率計算において，ベクトルの次元の差異がそのベクトルの生起確率の高低に影響し，次元数の大きいベクトルが軽視されてしまい易いという問題点を解決する学習アルゴリズムを提案する．

3.2　変数の定義

J 個のアピール集合を $\mathcal{A}=\{a_j : 1 \le j \le J\}$，$I$ 個の理由集合を $\mathcal{R}=\{r_i : 1 \le i \le I\}$ と定義する．全 N 件の理由登録のうち，n 番目に理由登録された企業のアピールベクトル集合を $\mathcal{X}=\{\boldsymbol{x}_n : 1 \le n \le N\}$，学生の理由ベクトル集合を $\mathcal{Y}=\{\boldsymbol{y}_n : 1 \le n \le N\}$ とし，n 番目に理由登録された企業のアピールベクトルを $\boldsymbol{x}_n=(x_1^n, x_2^n, ..., x_J^n)$，$n$ 番目の理由登録に対する理由ベクトルを $\boldsymbol{y}_n=(y_1^n, y_2^n, ..., y_I^n)$ と表す．ただし，x_j^n, y_i^n はそれぞれ n 番目の理由登録でアピール a_j，理由 r_i が選択されていれば 1，そうでなければ 0 を取る二値変数である．また，K 個の潜在クラス集合を $\mathcal{Z}=\{z_k : 1 \le k \le K\}$ と定義する．

3.3　モデルの提案

本研究では，各理由登録ごとに潜在クラスを仮定することで，アピールと理由の関係性を抽出することを目的とし，以下に示す確率モデルを提案する．これにより，各潜在クラスは項目ごとの関係性を表現することができる．

ここで，n 番目の理由登録に注目し，所属する潜在クラスを $v_n \in \mathcal{Z}$ とし，n 番目の理由登録を表すデータを $(\boldsymbol{x}_n, \boldsymbol{y}_n, v_n)$ と表すものとする．各アピール，各理由の生起確率モデルにはそれぞれ二項分布を仮定する．このとき，v_n を含む完全データの確率モデル $P(\boldsymbol{x}_n, \boldsymbol{y}_n, v_n)$ は式 (1) で表される．

$$\begin{aligned} P(\boldsymbol{x}_n, \boldsymbol{y}_n, v_n) &= P(v_n)P(\boldsymbol{x}_n|v_n)P(\boldsymbol{y}_n|v_n) \\ &= P(v_n)\prod_{j=1}^{J} P(a_j|v_n)^{x_j^n} P(\overline{a_j}|v_n)^{1-x_j^n} \\ &\quad \times \prod_{i=1}^{I} P(r_i|v_n)^{y_i^n} P(\overline{r_i}|v_n)^{1-y_i^n} \quad (1) \end{aligned}$$

ただし，$\overline{a_j}$，$\overline{r_i}$ はアピール a_j, 理由 r_i を選択しない事象で，$P(a_j|v_n)+P(\overline{a_j}|v_n)=1$,$P(r_i|v_n)+P(\overline{r_i}|v_n)=1$ を満たす．

式 (1) に対する対数尤度関数 LL は次式で与えられる．

$$LL = \sum_{n=1}^{N}\sum_{v_n \in \mathcal{Z}} \log P(\boldsymbol{x}_n, \boldsymbol{y}_n, v_n) \quad (2)$$

表 1: 各アピールの生起確率の比較

	企業理念	ビジョン	事業の特徴	事業優位性	仕事内容	魅力的人材	社風	制度	職場環境	技術力	採用方針
$P(a_j\|z_1)/P(a_j)$	**1.07**	0.74	**1.14**	**1.07**	**1.29**	**1.49**	0.98	0.91	0.62	0.02	1.00
$P(a_j\|z_2)/P(a_j)$	**1.84**	0.64	0.62	0.46	0.27	0.76	**1.30**	0.36	**3.47**	**2.92**	0.35
$P(a_j\|z_3)/P(a_j)$	0.57	**1.18**	0.17	**1.48**	**1.34**	**1.42**	0.75	**1.21**	0.46	**1.59**	**2.43**
$P(a_j\|z_4)/P(a_j)$	**1.20**	**1.44**	0.84	0.35	0.03	**1.21**	**1.31**	**2.05**	**2.50**	0.39	**1.12**
$P(a_j\|z_5)/P(a_j)$	0.80	**1.22**	**1.46**	1.00	0.97	0.27	0.98	0.98	0.48	**1.22**	0.44

表 2: 各理由の生起確率の比較

	仕事内容	業界	勤務地	職場雰囲気	ビジョン	ステータス	経験の活用	商品の内容	事業戦略	事業成長性
$P(r_i\|z_1)/P(r_i)$	0.95	**1.13**	**1.11**	**1.17**	**1.14**	0.78	0.88	0.81	0.47	0.50
$P(r_i\|z_2)/P(r_i)$	0.74	**1.04**	**1.06**	0.79	0.99	0.88	0.40	0.89	0.93	**1.01**
$P(r_i\|z_3)/P(r_i)$	**1.07**	**1.05**	0.99	0.13	0.07	0.66	**1.06**	0.66	0.28	0.39
$P(r_i\|z_4)/P(r_i)$	**1.09**	0.31	0.95	**2.11**	**1.38**	0.68	0.93	0.65	**1.45**	**1.24**
$P(r_i\|z_5)/P(r_i)$	**1.09**	0.96	0.87	**1.09**	**1.26**	**1.54**	**1.34**	**1.51**	**1.90**	**1.82**

	給与水準	職場環境	福利厚生	勤務形態	教育制度	能力主義	昇進の早さ	自己成長性	採用難易度	社会貢献度
$P(r_i\|z_1)/P(r_i)$	0.60	0.78	0.71	0.79	0.75	0.25	0.44	0.64	0.96	0.86
$P(r_i\|z_2)/P(r_i)$	0.19	0.83	0.34	0.22	0.52	0.83	0.56	0.94	0.87	0.70
$P(r_i\|z_3)/P(r_i)$	0.74	0.28	0.62	0.60	0.54	0.32	0.44	0.74	0.74	0.16
$P(r_i\|z_4)/P(r_i)$	**1.56**	0.72	**1.91**	**1.95**	**2.87**	**2.99**	**2.83**	**1.32**	**1.68**	**1.25**
$P(r_i\|z_5)/P(r_i)$	**1.75**	**1.76**	**1.56**	**1.52**	**1.27**	**1.81**	**1.67**	**1.49**	**1.07**	**1.66**

3.4 EM アルゴリズムによるパラメータの推定

提案モデルにおけるパラメータ $P(z_k)$,$P(a_j|z_k)$,$P(r_i|z_k)$ は EM アルゴリズム [3] により推定を行う．このアルゴリズムでは，式 (2) が収束するまで E-step，M-step を繰り返し計算し，パラメータを推定する．各 Step の更新式は以下の式 (3), 式 (4)-(6) となる．

E-step)

$$P(z_k|\boldsymbol{x}_n,\boldsymbol{y}_n) = \frac{P(\boldsymbol{x}_n,\boldsymbol{y}_n,z_k)}{\sum_{z_k\in\mathcal{Z}}P(\boldsymbol{x}_n,\boldsymbol{y}_n,z_k)} \quad (3)$$

M-step)

$$P(z_k) = \frac{1}{N}\sum_{n=1}^{N}P(z_k|\boldsymbol{x}_n,\boldsymbol{y}_n) \quad (4)$$

$$P(a_j|z_k) = \frac{1}{NP(z_k)}\sum_{n=1}^{N}P(z_k|\boldsymbol{x}_n,\boldsymbol{y}_n)x_j^n \quad (5)$$

$$P(r_i|z_k) = \frac{1}{NP(z_k)}\sum_{n=1}^{N}P(z_k|\boldsymbol{x}_n,\boldsymbol{y}_n)y_i^n \quad (6)$$

本研究で対象となるアピールと理由は，各理由登録に対してそれぞれベクトルで与えられ，式 (1) における $P(\boldsymbol{x}_n|v_n)$, $P(\boldsymbol{y}_n|v_n)$ はベクトル内の各要素の二項分布の積の形で計算される．そのため，$I\neq J$ のとき，要素数が多い方のベクトルの生起確率は低くなりやすく，結果として要素数が少ない方のベクトルの特徴に過剰に適合した潜在クラスを構成してしまうという問題が存在する．この問題を解決するために，式 (3) における $P(\boldsymbol{x}_n,\boldsymbol{y}_n,z_k)$ を次式で置き換えた学習アルゴリズムを提案する．

$$P'(\boldsymbol{x}_n,\boldsymbol{y}_n,v_n) = P(v_n)P(\boldsymbol{x}_n|v_n)^{\alpha}P(\boldsymbol{y}_n|v_n)^{\beta} \quad (7)$$

ただし，α，β は重みパラメータであり，ベクトルの次元数や各要素の確率分布を考慮して事前に決定する必要がある．

4 分析

4.1 分析概要

提案手法を用いて，サイト A における 2016 年卒業学生を対象としたアピールデータおよび理由データを分析する．アピール数および理由数は $J=11$，$I=20$ である．また，分析対象の企業として，アピールを 3 つ掲載している，従業員規模 1,000 人未満の企業に限定した．学生の理由登録データは，2014 年 6 月から 2015 年 9 月までの期間における対象企業への全理由登録データからランダムに 100,000 件を抽出した．また，事前実験の結果より，潜在クラス数 $K=5$，式 (7) における重みパラメータは $\alpha=1.25$，$\beta=1.00$ とした．

4.2 分析結果と考察

提案手法の適用により得られた各アピール，各理由の各潜在クラスのもとでの生起確率 $P(a_j|z_k)$，$P(r_i|z_k)$ を，基準となる $P(a_j)$，$P(r_i)$ との比で表したものを表 1，表 2 に示す．ただし，それぞれの値が 1 を超える場合に太字にした．

表 1，表 2 を併せて比較検討することで，各アピールと各理由の関係性を分析する．潜在クラス z_2,z_4 に着目すると，表 1 より,「社風」,「職場環境」のアピールの生起確率がそれぞれ高い．表 2 の z_4 では「職場雰囲気」,「勤務形態」の理由の生起確率が高い．一方で表 2 の z_2 では,「業界」,「勤務地」の生起確率が高いが，それ以外の生起確率は低い．よって，z_4 は「アピールと理由がマッチしているクラス」，z_2 は「アピールが理由に影響をあまり与えていないクラス」と解釈できる．すなわち，z_2 に所属する志望理由登録をした学生は「業界」や「勤務地」を企業志望の軸としていると考えられる．次に潜在クラス z_5 に着目する．表 1 では,「ビジョン」,「事業の特徴」のアピールの生起確率が高い．表 2 では,「ビジョン」,「商品の内容」の理由の生起確率が高く，その他の理由の生起確率も全体的に高い．よって，z_5 は「アピールと理由がマッチしている部分に加え，アピールとは異なる理由も選択されているクラス」と解釈できる．以上のように，提案モデルによりアピールと理由の関係性を定量的に分析することが可能となった．

5 まとめと今後の課題

本研究では，アピールと理由の関係性を分析するための潜在クラスモデルを新たに提案した．また，提案したモデルを実データに適用し，関係性の分析を行った．

今後の課題として，得られた結果の要因と考えられるアピール本文や，依存している学生や企業の情報を分析する必要があり，それらを考慮したモデルの拡張が望まれる．

参考文献

[1] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell,“Text Classification from Labeled and Unlabeled Documents using EM,” *Machine Learning*, Vol.39, pp.103-134, 2000.

[2] T. Hofmann,“Probabilistic Latent Semantic Analysis,” *Proc.of UAI’99*, pp.289-296, 1999.

[3] 宮川雅巳,“EM アルゴリズムとその周辺,” 応用統計学, Vol.16, No.1, pp.1-21, 1987.