就職ポータルサイトにおけるアスペクトモデルを用いた推薦モデルに関する研究

情報数理応用研究　　5213C034-5　古山亮
指導教員　後藤正幸

A Study of Recommender System on Internet Portal Sites for Job Hunting Using Aspect Model

FURUYAMA Ryo

1　研究背景・目的

近年，多くの学生がインターネット上の就職ポータルサイトを用いて就職活動を行なうようになった．就職ポータルサイトは，学生を採用する立場にある多くの企業にとって，自社選考への申込を受付ける場であり，学生は就職ポータルサイトを通じて選考参加への意思表示（以下，エントリと呼ぶ）を行うことができる．学生は就職ポータルサイトを通じて様々な企業にもエントリを行える．一方，就職ポータルサイトには数多くの企業が掲載されているが，学生のエントリは一部企業に集中しがちであり，学生・企業間のミスマッチが問題視されている．

このような問題に対処するため，多くの就職ポータルサイトに，学生がエントリする確率の高いであろう企業を推薦する推薦システムが導入されている．推薦システムは EC サイトなどにおいて，特定のモデルまたはロジックに基づき，ユーザに推薦するアイテムを決定し，ダイレクトメールや関連商品ページなどでユーザにアイテムを推薦するシステムである．

一般に，EC サイトなどにおける推薦システムでは潜在クラスモデルを用いた手法の有効性が認められている [1]．しかし，就職ポータルサイトには，「広報開始月にユーザが一斉にエントリ行動を開始する」，「年度ごとにユーザが殆ど入れ替わる」，「推薦するアイテムの増減はあまり見られない」といった，一般的な EC サイトには見られない特性が存在する．そのため，潜在クラスモデルを直接用いる推薦システムが就職ポータルサイトでも同様に有効かどうかという点については疑問が残る．

そこで本研究では，これら就職ポータルサイトの特性に対応した潜在クラスモデルに基づく用いた推薦システムを提案する．具体的には，ユーザのエントリ履歴の蓄積が浅いサービス序盤において，前年度以前のエントリ履歴を学習した潜在クラスモデルを活用することで，推薦精度の向上を図る手法を提案する．また，提案手法の有効性を検証するため，大手就職ポータルサイト（以下，サイト A と呼ぶ）における 2013 年卒のエントリデータを学習データ，2014 年卒のエントリデータをテストデータとしてシミュレーション実験を行う．

2　準備

2.1　就職ポータルサイト

就職ポータルサイトとは，包括的に学生の就職活動を支援する Web サービスである．近年の爆発的な情報技術の普及に伴い，学生の就職活動は，大学の就職課や学生の元へ送られる求人広告などを媒体とした従来の方法から，Web サイトや電子メール等といった情報技術を活用した方法へと大きく変化している．就職ポータルサイト上でのみエントリの受け付けや採用情報の掲載を行っている企業も少なくなく，就職ポータルサイトを経由したエントリを行う方法が企業・学生間共に主流となっている．

2.2　就職ポータルサイトを用いた就職活動の流れ

学生は就職ポータルサイトに掲載されている企業の中から自らの求める条件に合致する企業を検索する．この際，個社ページに掲載されている企業情報や企業紹介文を参考にする．志望企業を絞込んだのち，エントリすることで，各企業が設ける入社試験に臨む．

このように，学生は就職ポータルサイトを利用する過程で様々な企業と接点を持つ．したがって，学生にとって就職ポータルサイトは単にエントリを行う場所というわけではなく，今までに知ることのなかった企業を新たに見つける場所でもある．また採用活動を行う企業側にとっても，就職ポータルサイトは学生に自社を知ってもらうための有用な広報の窓口となる．

2.3　就職ポータルサイトにおける推薦モデル

前述のように，学生は就職ポータルサイト上で興味のある企業を探す段階で，業種や従業員規模，所在地といった条件検索や，フリーワード検索などを行う．しかし，サイト A には数多くの企業が登録されており，その中から興味に合致した企業を数多く探しだすことは難しく，学生が本来興味を持つであろう企業を見逃している可能性がある．そこで，サイト運営者はサイト上での学生のエントリを促すため，興味のある企業を学生に認知させる推薦システムを導入している．

推薦システムは EC サイトなどの Web サービスに広く導入されているシステムで，ユーザの嗜好に合致するであろうアイテムを推薦することでサービスの利便性を向上させる狙いがある．就職ポータルサイトにおける推薦システムは，特定の推薦モデルに基づいて学生へ推薦する企業を決定し，ダイレクトメールなどを用いて学生に適切な企業を推薦するシステムである．ここで，一般的な EC サイトにおける商品が企業に，ユーザが学生に対応している．

サイト A における推薦は，ある推薦モデルに基づき，各学生がエントリする企業を予測している．そして，予測された企業の企業名・業種・本社所在地・事業概要などを記載したリストを一定数，各学生に提示している．このため，どの企業を推薦すべきかを決定する推薦モデルは推薦システムにおいて，非常に重要な要素となっている．

就職ポータルサイトには学生の行動履歴が蓄積されている．そのため，これらの大量のデータを推薦に活用す

ることで，より精度の高い推薦を行うことが可能であると考えられる．

3　従来手法（一般的な潜在クラスモデル）

本項では，一般的な潜在クラスモデルのひとつである Aspect Model[1]-[3]（以下，AM と呼ぶ）について述べる．AM は，学生のエントリ傾向および企業の被エントリ傾向を推定するために用いることのできる確率的潜在クラスモデルであり，文書分析や EC サイトにおける協調フィルタリングなどに応用されている．このモデルでは，学生と企業の間に潜在クラスが仮定されており，類似した傾向を持つ学生および類似した被エントリ傾向を持つアイテムは同じ潜在クラスに属するものとの仮定をおいている．

また，このモデルにおいて学生と企業は単一の潜在クラスでなく，複数の異なる潜在クラスに所属できることを仮定している．この 2 つの仮定により，このモデルでは学生のエントリ傾向と企業の被エントリ傾向の多様な表現が可能となる．AM のグラフィカルモデルは図 1 で示される．

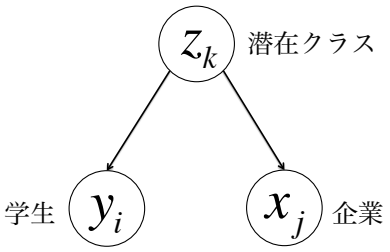

図 1: AspectModel のグラフィカルモデル

いま，I 人の学生集合を $\mathcal{Y} = \{y_i : 1 \leq i \leq I\}$，$J$ 社の企業集合を $\mathcal{X} = \{x_j : 1 \leq j \leq J\}$，$K$ 個の潜在クラス集合を $\mathcal{Z} = \{z_k : 1 \leq k \leq K\}$ とする．このとき，学生 y_i が企業 x_j にエントリするという事象を (y_i, x_j) と定義する．この時，AM のモデルは以下の式 (1) のように表すことができる．

$$P(y_i, x_j) = \sum_{k=1}^{K} P(z_k)P(y_i \mid z_k)P(x_j \mid z_k) \tag{1}$$

ここで，それぞれのパラメータ $P(z_k)$，$P(y_i \mid z_k)$，$P(x_j \mid z_k)$ は，EM アルゴリズムを用い，以下の式 (2) における対数尤度関数 LL を最大化することで推定できる．

$$LL = \sum_{i=1}^{I}\sum_{j=1}^{J} \delta(y_i, x_j)\log P(y_i, x_j) \tag{2}$$

$\delta(y_i, x_j)$ は，学生 y_i が企業 x_j にエントリしていた場合 1，そうでない場合 0 を返す指示関数である．パラメータの推定後，推定されたパラメータを以下の式 (3) のように用いることによって，ユーザ y_i の未エントリ企業 x_j に対するエントリ確率を算出できる．

$$\hat{P}(x_j|y_i) = \frac{\sum_{k=1}^{K} \hat{P}(y_i|z_k)\hat{P}(x_j|z_k)\hat{P}(z_k)}{\sum_{j=1}^{J}\sum_{k=1}^{K} \hat{P}(y_i|z_k)\hat{P}(x_j|z_k)\hat{P}(z_k)} \tag{3}$$

ここで，$\hat{P}(z_k)$，$\hat{P}(y_i \mid z_k)$，$\hat{P}(x_j \mid z_k)$ は EM アルゴリズムを用いて推定されたパラメータである．式 (3) で示されたエントリ確率の高い企業から順に推薦を行うことで，より適切な企業をエントリするよう学生に促すことができる．

4　提案手法

4.1　本研究のアプローチ

従来の AM は，一般的な EC サイトでの推薦システムにおいて有用性が認められている．一方で，就職ポータルサイトにおいては前述の通り「広報開始月に学生が一斉にエントリ行動を開始する」,「卒業年度ごとに学生がほとんど全て入れ替わる」,「推薦対象企業の増減はほとんど見られない」，といった一般的な EC サイトには見られない独特の特色がある．

行動履歴を持たない新規ユーザへの推薦が行えない，いわゆるコールドスタート問題に対応した手法としては，ユーザ情報を利用して推薦を行う Xuan ら [4] の手法などが挙げられる．しかし，同様のサービスが周期的に提供される就職ポータルサイトにおいては，各年度でユーザが殆ど入れ替わるうえ，年齢や職業も一定となってしまうため，適用が難しい．

また，従来の AM では学生集合 $\mathcal{Y}$ には変化がないことが仮定されているが，この仮定の下では，就職ポータルサイトにおける推薦システムは単年度ごとに独立したものとなってしまう．しかしながら，年度が変化しても就職活動を行う学生の行動傾向の周期性は保たれるものと考えられる．就職ポータルサイトには過去の年度のエントリデータも蓄積されており，これらを活用することで，単年度のデータのみでは予測の行えない，季節ごとの学生の行動傾向の変化なども考慮に入れた有効性の高い推薦が行える可能性がある（図 2）．

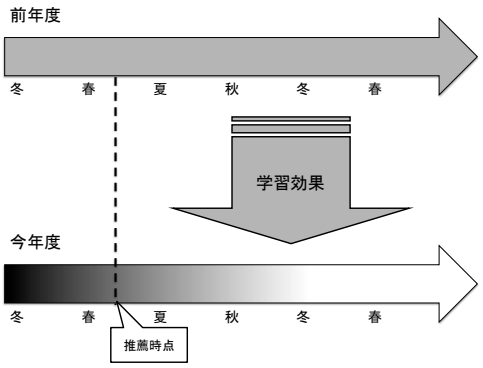

図 2: 本研究のアプローチ

したがって，本節では年度が変わり，就職活動の比較的初期段階の学生に対して推薦を行う AM を提案する．具体的には，前年度以前のエントリデータでモデルの学習を行い，学習されたモデルとは異なる学生集合のエントリ確率を算出する手法を提案する．ただし，年度を跨いだ場合でも企業集合 $\mathcal{X} = \{x_j : 1 \leq j \leq J\}$ には変化がないことを仮定している．

4.2 提案手法 1

提案手法 1 では，k-NN 法のアイディアを取り入れ推薦対象の学生と類似する複数の学生を前年度の学習データから選定し，その多数決を取る方法を考える．すなわち，学習データの中で最も類似している N 人の学生を選び，そのエントリ傾向から推薦すべき企業を決定する．

学習されたモデルにおける学生集合 $\mathcal{Y}$ に対し，現在の学生集合を $\mathcal{Y}^* = \{y_l^* : 1 \leq l \leq L\}$ と定義する．提案手法 1 のイメージを図 3 に示す．

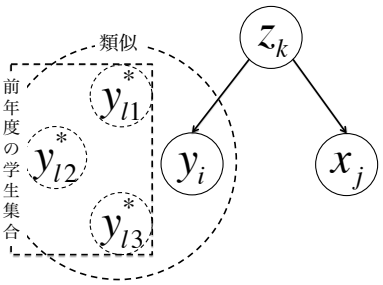

図 3: 提案手法 1 のイメージ

また，y_l^* に対応する前年度以前の N 人の類似した学生集合を，$\mathcal{S}(y_l^*) = \{y_{l1}, ..., y_{ln}, ..., y_{lN}\}$ とする．類似度は次の式 (4) で計算する．式 (4) は今年度の学生 y_l^* と前年度の学生 y_i の間で同じ企業へのエントリが行われた数を示している．

$$sim(y_i, y_l^*) = \sum_{j=1}^{J} \delta(y_i, x_j)\delta(y_l^*, x_j) \tag{4}$$

対応学生を決定したのち，従来と同様に以下の式 (5) でエントリ確率の算出を行う．

$$\hat{P}(x_j|y_l^*) = \frac{1}{|\mathcal{S}(y_l^*)|} \sum_{y_{ln} \in \mathcal{S}(y_l^*)} \frac{\alpha_j}{\sum_{j=1}^{J} \alpha_j} \tag{5}$$

ただし

$$\alpha_j = \sum_{k=1}^{K} \hat{P}(y_{ln}|z_k)\hat{P}(x_j|z_k)\hat{P}(z_k)$$

である．

式 (5) は，各対応学生が企業 x_j にエントリを行う確率の平均となっている．式 (5) の高い企業から順に推薦を行う．

4.3 提案手法 2

前述した提案手法 1 では，学習モデルにエントリ傾向の似た学生が存在することを暗黙のうちに仮定している．そのため本項では前年度以前の特定の学生に依存せず，前年度以前に学習されたパラメータを直接利用してエントリ確率を算出する方法を提案する．学生 y_l^* が潜在クラス z_k に所属する確率を $\hat{P}(z_k|y_l^*)$ とするとき，$\hat{P}(z_k|y_l^*)$ は，エントリを行った企業 x_j の潜在クラス z_k に対する所属確率 $\hat{P}(z_k|x_j)$ を用いて，次の形で表すことができる．

$$\hat{P}(z_k|y_l^*) = \frac{\prod_{j=1}^{J} \delta(y_l^*, x_j)\hat{P}(z_k|x_j)}{\sum_{k=1}^{K}\prod_{j=1}^{J} \delta(y_l^*, x_j)\hat{P}(z_k|x_j)} \tag{6}$$

[1]12 月末まで，1 月末まで…8 月末までと 1ヶ月刻みで変化

ここでベイズの定理 [5] より，$\hat{P}(z_k|x_j)$ は，学習したパラメータ $\hat{P}(x_j|z_k)$ を利用して，次の形で表すことができる．

$$\hat{P}(z_k|x_j) = \frac{\hat{P}(x_j|z_k)\hat{P}(z_k)}{\hat{P}(x_j)} \tag{7}$$

式 (6) および式 (7) より，最終的に $\hat{P}(z_k|y_l^*)$ は次の式 (8) で表すことができる．

$$\hat{P}(z_k|y_l^*) = \frac{\prod_{j=1}^{J} \delta(y_l^*, x_j)\hat{P}(x_j|z_k)\hat{P}(z_k)}{\sum_{k=1}^{K}\prod_{j=1}^{J} \delta(y_l^*, x_j)\hat{P}(x_j|z_k)\hat{P}(z_k)} \tag{8}$$

このとき，エントリ確率は以下の式 (9) で算出する．

$$\hat{P}(x_j|y_l^*) = \sum_{k=1}^{K} \hat{P}(z_k|y_l^*)\hat{P}(x_j|z_k) \tag{9}$$

こののち，従来手法や提案手法 1 と同様に，式 (9) より得られたエントリ確率の高い企業から推薦を行う．

5 実験

本項では，実際にサイト A に蓄積されたデータを用いてシミュレーションを行い，この結果から提案手法の有効性を検討する．

5.1 実験条件

実験はサイト A の 2013 年卒学生のエントリデータで学習を行い，2014 年卒学生のエントリデータをランダムサンプリングしてテストを行った．学習データ及びテストデータの詳細は以下のとおりである．

学習データ: 2011 年 12 月〜2013 年 3 月（約 200 万件）

テストデータ: 2012 年 12 月〜一定期間（約 180 万件）[1]

推薦対象の学生数: 2000 人
（テストデータよりサンプリング）

潜在クラス数: 10, 20, 30

N（提案手法 1）: 1,2,3 と変化

評価指標には Top10 精度を用いる．これは，推薦候補として上がった 10 件の企業のうち，どれだけ実際にエントリされていたかを示す指標である．

5.2 結果・考察

各手法において最も高い精度を示した，潜在クラス数 $K = 10$ で実験を行った場合の Top10 精度を図 4 に示す．

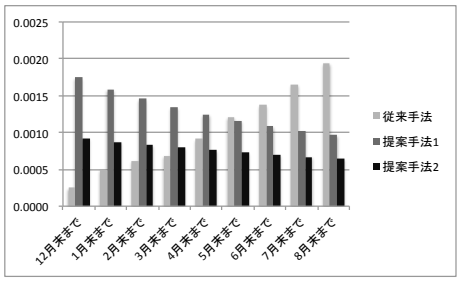

図 4: 実験結果（$K = 10$）

提案手法 1, 2 ともに新年度が始まるまでのサービス序盤では従来手法よりも高い精度での推薦を行えることが確認できた．従来手法は期間が経つにつれ精度が向上した．従来手法はテストに用いる期間を増やしていくごとにエントリデータが蓄積されていくため，推薦の精度が向上していったものと考えられる．提案手法 1, 2 はともに期間が経つにつれ精度が低下していく傾向にあった．提案手法 1 および提案手法 2 は序盤では高い精度を保っているものの，期間が経つにつれ精度が低下していった．これはエントリデータの蓄積に伴って，現在の企業の潜在クラスへの所属確率と，前年度に算出したものとのずれが大きくなっていったためと考えられる．また，潜在クラス数を $K = 20, 30$ とした場合の Top10 精度も図 5，図 6 に示す．

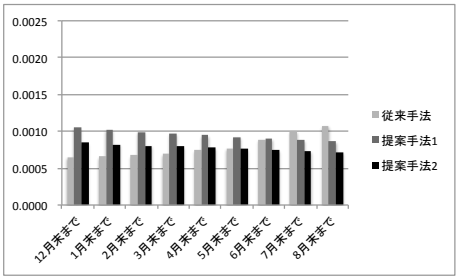

図 5: 実験結果（$K = 20$）

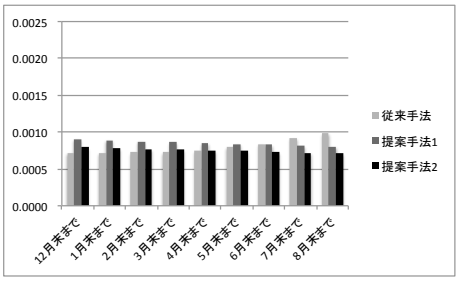

図 6: 実験結果（$K = 30$）

潜在クラス数を増やしていった結果，精度は低下していき，また各手法で精度の差が小さくなっていくことが分かった．

また，提案手法 1 について，N を変化させた場合の Top10 精度を図 7 に示す．

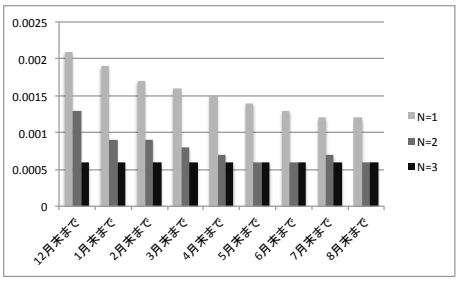

図 7: 実験結果（提案手法 1）

N が増えていくごとに精度は低下していく結果となった．この理由として，潜在クラスへの所属確率の推定によって学生の分類と同様の効果が生まれているため，エントリ傾向の似た学生を複数考慮してもあまり効果的でないことが考えられる．また特徴量となる企業数が大きいため，単純に平均をとると各企業へのエントリ確率がばらつきやすいといった理由も考えられる．

また，提案手法で用いた潜在クラスモデルが実際に有効にあてはまり，なんらかの特徴群に分かれていることを確認するため，各潜在クラスを分析した．サイト A において設定されている業種・所在地・従業員規模といった特徴に着目した．分析の結果，各潜在クラスは表 1 のような特徴を持つことが分かった．

第 3 次産業とされる業種が多く見られた．なかでも情報処理・ソフトウェア産業の企業へのエントリがいずれの潜在クラスでも多く見られたが，これはそもそもそれら業種に属する企業が多いことに起因すると考えられるため，ここでは除外した．

表 1: 各潜在クラスの企業の特徴 (抜粋)

潜在クラス	特徴
1	クレジット信販
2	商社（医療機器）
3	服飾雑貨・皮革製品
4	商社（繊維製品）
5	文具・事務機器・インテリア
6	商社（アパレル・服飾雑貨・貴金属）・百貨店
7	専門店（複合）
8	商社（化粧品）
9	安全・セキュリティ産業
10	エステ・理容・美容

また，所在地・従業員規模に関しては特徴的な傾向は見られなかった．この結果，本研究で適用した潜在クラスモデルにより，企業の業種に特徴を持つクラスに上手くクラスタリングされていることを確認できる．

6 まとめと今後の課題

本研究では，就職ポータルサイトにおけるユーザの入れ替わりという特色に対応し，潜在クラスモデルを用いた推薦手法を提案した．また，シミュレーション実験により，実際に異なる学生集合から学習したモデルを用いて推薦が行えることを示し，サービス序盤で従来手法よりも高い精度で推薦を行えることを確認した．

モデルの拡張として，前年度以前のデータと推薦対象年度のデータを共に用いたモデルや，学生のエントリ傾向の経時変化を考慮したオンライン学習モデルを提案することで，更なる推薦精度の向上が見込めるものと考える．また，関連した問題の学習結果を再利用する転移学習 [6] の考えを取り入れることや，LDA[7] との比較などが課題として挙げられる．

参考文献

[1] T.Hofmann and J.Puzicha, “Latent Class Models for Collaborative Filtering”, *Proc. 16th International Joint Conference on Artificial Intelligence*, pp.688–693, 1999.

[2] T.Hofmann, “Probabilistic Latent Semantic Analysis”, *UAI*, pp.289–296, 1999.

[3] T.Hofmann, “Latent Semantic Models for Collaborative Filtering”, *ACM Transactions on Information Systems*, Vol.22, No.1, pp.89–115, 2004.

[4] X.N.Lam, T.Vu, T.D.Le and A.D.Duong, “Addressing cold-start problem in recommendation system”, *ICUIMC ’08 Proceedings of the 2nd international conference on Ubiquitous information management and communication*, pp.208–211, 2008.

[5] 後藤正幸, 小林学, “入門 パターン認識と機械学習”, コロナ社, pp.184, 2014.

[6] T.Kamishima, “Transfer Learning”, *Journal of Japanese Society for Artificial Intelligene 25(4)*, pp.572–580, 2010.

[7] D.Blei, A.Ng and M.Jordan, “Latent Dirichlet Allocation”, *Journal of Machine Learning Research*, pp.1107–1135, 2003.