就職ポータルサイトにおける嗜好の時間的変化を考慮したユーザクラスタリング手法の提案

1X11C082-0　永森 誠矢
指導教員　後藤 正幸

1　研究背景・目的

近年，学生 (以下ユーザ) の就職活動において就職ポータルサイトの利用が一般的となっている．これにより，ユーザは多くの企業へ容易にエントリが可能である一方，ユーザと企業のミスマッチによる就職活動長期化が問題となっている．この問題の改善のため，就職ポータルサイトに蓄積されるユーザの行動履歴データを活用することで，将来的に就職活動終了時期が遅くなるであろうユーザを早期に発見し，サポートできる可能性がある．

ユーザには 1 つの業種を粘り強くエントリするユーザや，就職活動を通じて自分の適性などを評価しエントリ行動を変化させていくユーザなどがいる．このようなユーザの就職活動を通じて行われるエントリ傾向の時間的変化は，就職活動終了時期に影響すると考えられる．しかし，エントリ行動から就職活動終了時期の予測モデルを構築することを考える際，ユーザの就職活動終了時期にはエントリ行動の他，様々な外乱要因による影響が存在し，決定木のような通常の予測モデルを用いた方法では精度のよいモデルを構築することが困難になってしまう．

そこで本研究では大局的な観点での就職活動終了時期の予測モデルの構築を試みる．具体的には，時間的変化を考慮したエントリ傾向に基づくクラスタリングを行うことにより，各クラスタのエントリ傾向の分析が可能であり，各クラスタの就職活動終了時期の早遅を判別する予測モデルを提案する．このクラスタリングにより大局的な観点から説明可能な就職活動終了時期の予測モデルを構築し，将来的に就職活動終了時期が遅くなると想定されるユーザを予測することができる．提案手法の有効性を示すために就職ポータルサイトのデータを用いたシミュレーション実験を行う．加えて，提案手法を用いた知識発見の方法として，形成したクラスタに所属するユーザの特徴を分析可能であることを示す．

2　Aspect Model [1]

本研究ではユーザのエントリ傾向を定量化するために Aspect Model(以下 AM) を適用する．AM はユーザと企業の間に潜在クラスを仮定し，ユーザと企業を確率的にクラスタリングする統計モデルである．いま，I 社の企業集合を $\mathcal{X} = \{x_i : 1 \leq i \leq I\}$，$J$ 人のユーザ集合を $\mathcal{Y} = \{y_j : 1 \leq j \leq J\}$，$K$ 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，AM の確率モデルは以下の式 (1) で示される．

$$P(x_i, y_j) = \sum_k P(z_k)P(x_i|z_k)P(y_j|z_k) \tag{1}$$

ただし，式 (1) における $P(z_k)$，$P(x_i|z_k)$，$P(y_j|z_k)$ は EM アルゴリズムにより推定する．

3　提案手法

本研究では就職活動終了時期が遅くなると想定されるユーザを予測するため，時間的変化を考慮したユーザのエントリ傾向のパターンと就職活動終了時期の関係をモデル化すると共に，それらのエントリ傾向のパターンを分析可能なモデルを構築する．しかし，企業数は非常に多いため，個々の企業へのエントリデータを集約し，類似した企業を一括して扱う必要がある．そこで，潜在クラスモデルである AM を用いることにより，ユーザのエントリ傾向を定量化し，これらをクラスタリングするモデルを考える．AM の潜在クラスへの所属確率を用いれば，エントリ傾向が定量化できると共に，その時間推移も容易に計算可能である．

本研究の提案手法の手順は，AM によるユーザのエントリ傾向の定量化，k-means 法による学習ユーザのクラスタリング，クラスタとの類似度計算による予測対象ユーザの所属クラスタの推定と就職活動終了時期の予測からなる．提案であるクラスタリング手法を以下のアルゴリズムで行う．

Step1 AM による学習ユーザの時期ごとのエントリ傾向の定量化
Step2 k-means 法による学習ユーザのクラスタリング
Step3 AM による予測対象ユーザの時期ごとのエントリ傾向の定量化
Step4 類似度計算による予測対象ユーザの所属クラスタの推定
Step5 所属クラスタによる就職活動終了時期の予測　□

ここで，学習ユーザは，AM を構築し，就職活動終了時期の予測に用いるクラスタ形成のためのユーザを表す．

3.1　ユーザのエントリ傾向の定量化

いま，ユーザは就職活動を行う際に企業に対し嗜好を持ち，ユーザの企業への嗜好は潜在クラスへの所属確率により表現可能であるとする．エントリ傾向の時間的変化を考慮するため，エントリ傾向を T 期間に分けて算出することとし，ユーザ y_j の t 期のエントリ傾向を式 (2) により定量化する．

$$P'_t(z_k|y_j) = \frac{1}{N_{jt}} \sum_i \eta_{ijt} \hat{P}(z_k|x_i) \tag{2}$$

ただし，η_{ijt} は t 期にユーザ y_j が企業 x_i にエントリしているときに 1 を，それ以外は 0 をとる指示関数とし，$t(t$=1,2,$\cdots$,$T)$ は時期，N_{jt} はエントリ傾向を求める時点 t 期でのユーザ y_j の総エントリ数を表している．また，各企業の潜在クラスへの所属確率 $\hat{P}(z_k|x_i)$ は，推定された AM のパラメータ $\hat{P}(z_k)$，$\hat{P}(x_i|z_k)$ を用いて以下の式 (3) により求める．

$$\hat{P}(z_k|x_i) = \frac{\hat{P}(z_k)\hat{P}(x_i|z_k)}{\sum_k \hat{P}(z_k)\hat{P}(x_i|z_k)} \tag{3}$$

ユーザは各潜在クラスに確率的に嗜好を持つことになり，各潜在クラスへの嗜好の和は 1 となる．この嗜好全体で，エントリ傾向を表す．このエントリ傾向の算出を用いることで任意の時期のユーザのエントリ傾向を求めることができる．

3.2　k-means 法による学習ユーザクラスタリング

予測対象ユーザの就職活動終了時期を予測するためのクラスタを形成するため，式 (2) で求めた各期のユーザのエントリ傾向を特徴量とした学習ユーザのクラスタリングを行う．式 (2) で得られる T 期分のエントリ傾向をユーザ y_j の特徴量 $\boldsymbol{w}_j$ とし，クラスタリングを行う．ただし，特徴量 $\boldsymbol{w}_j$ は以下の式 (4) で表される．

$$\boldsymbol{w}_j = (\boldsymbol{s}_{j1}, \boldsymbol{s}_{j2}, \cdots, \boldsymbol{s}_{jT}) \tag{4}$$

ここで，$\boldsymbol{s}_{jt}$ は各期 t のユーザ y_j のエントリ傾向を表す K 次元のベクトルであり，次の式 (5) で表される．

$$\boldsymbol{s}_{jt} = (P'_t(z_1|y_j), P'_t(z_2|y_j), \cdots, P'_t(z_K|y_j)) \tag{5}$$

本研究ではクラスタリング手法として k-means 法を用いる．クラスタ数は C とし，ユーザとクラスタの類似度計算に用いるため，各クラスタの代表ベクトル $\boldsymbol{c}_l(l$=1,2,$\cdots$,$C)$ を以下の式 (6) で求める．

$$\boldsymbol{c}_l = \frac{1}{D_l} \sum_j q_{jl} \boldsymbol{w}_j \tag{6}$$

ただし，q_{jl} はユーザ y_j がクラスタ l に所属するときに 1 を，それ以外は 0 をとる指示関数とし，D_l はクラスタ l に属するユーザ数を表している．

3.3 予測対象ユーザの所属クラスタの推定

予測対象ユーザに対しても各期のエントリ傾向を算出する．企業は毎年ほぼ変わらないという就職ポータルサイトの特徴からユーザの潜在クラスへの嗜好は，式 (2) を直接用いて推定する．いま，M 人の予測対象ユーザの集合を $\mathcal{Y}' = \{y'_m : 1 \leq m \leq M\}$ とし，予測対象ユーザ y'_m に対して特徴量 $\boldsymbol{w}'_m$ を式 (7) で算出する．

$$\boldsymbol{w}'_m = (\boldsymbol{s}'_{m1}, \boldsymbol{s}'_{m2}, \cdots, \boldsymbol{s}'_{mT}) \tag{7}$$

ただし，$\boldsymbol{s}'_{mt} = (P'_t(z_1|y'_m), P'_t(z_2|y'_m), \cdots, P'_t(z_K|y'_m))$ とする．

次に予測対象ユーザの所属クラスタを求めるため，その特徴量 $\boldsymbol{w}'_m$ と，学習フェーズで形成された各クラスタの代表ベクトル $\boldsymbol{c}_l$ の類似度をユークリッド距離を用いて算出する．予測対象ユーザの所属クラスタは式 (8) により求めることとし，そのクラスタを $\hat{C}$ とする．

$$\hat{C} = \arg\min_l \|\boldsymbol{w}'_m - \boldsymbol{c}_l\|_2^2 \tag{8}$$

3.4 クラスタによる就職活動終了時期の予測

学習フェーズで形成されたクラスタごとに，所属する学習ユーザの平均就職活動終了日を計算しクラスタに付与する．予測対象ユーザの就職活動終了日は式 (8) で得られた $\hat{C}$ の平均値を予測値とする．

4 実験

エントリ傾向の時間的変化を考慮した提案手法が，就職活動終了時期が遅くなるユーザの予測に有効であることを示すため，実データを用いた実験を行った．加えて，提案手法の一応用として，得られたクラスタの分析を行い，そこに所属するユーザの特徴を把握することで有用な知見が得られることを示す．

4.1 実験条件

本実験では就職活動終了時期が遅くなると想定されるユーザの予測精度を比較するため，エントリ傾向の時間的変化を考慮した提案手法によるクラスタリングと時間的変化を考慮しないクラスタリングを用いる．

学習データとして 2013 年卒業のユーザ 141,434 人の全期間のエントリデータ 6,589,316 件，テストデータとして 2014 年卒業のユーザ 104,355 人の 3 月までのエントリデータ 4,864,984 件を用いた．就職活動終了時期が遅くなるユーザを発見することは早期であることが望ましい．そこで，本実験では 3 月までのエントリ傾向のみを用いて学習ユーザのクラスタリングを行い，予測対象ユーザの 3 月までのエントリ傾向から，その就職活動終了時期を予測する．また，潜在クラス数 K を K=10，クラスタ数 C を C=20，期間 T を T=3(12 月〜3 月 (t=1)，12 月〜1 月 (t=2)，2 月〜3 月 (t=3)) とした．比較手法として時間変化を考慮しない通常の AM を用いた．ただし，潜在クラス数 K は K=20 とし，予測対象ユーザの所属クラスの推定は，嗜好の最も高い潜在クラスにユーザが所属するものとした．

この実験では就職活動の終了時期が 9 月以降になるユーザを予測し，その精度を評価とすることとした．評価指標として再現率，精度，F 値を用いた．これら 3 つの指標は以下の式 (9)〜(11) で算出される．

$$\text{再現率} = \frac{\text{正しく予測したユーザ数}}{\text{就職活動終了が 9 月以降のユーザ数}} \tag{9}$$

$$\text{精度} = \frac{\text{正しく予測したユーザ数}}{\text{就職活動終了が 9 月以降と予測したユーザ数}} \tag{10}$$

$$F \text{ 値} = \frac{2 \times \text{再現率} \times \text{精度}}{\text{再現率} + \text{精度}} \tag{11}$$

比較手法における各潜在クラスの就職活動終了時期は，以下の式 (12) により予測する．

$$F(z_k) = \frac{\sum_j (P(z_k|y_j) \times U(y_j))}{\sum_j P(z_k|y_j)} \tag{12}$$

ただし，式 (12) における $F(z_k)$ は潜在クラス z_k の就職活動終了日，$U(y_j)$ はユーザ y_j の就職活動終了日である．

4.2 実験結果と考察

実験結果を図 1 に示す．

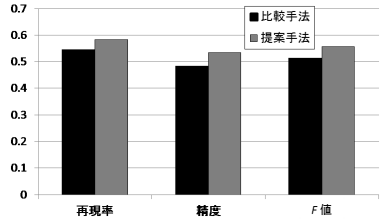

図 1. 再現率，精度，F 値の実験結果

図 1 より，提案手法は比較手法に比べ，再現率，精度，F 値において優れていることがわかる．提案手法では全体的なエントリ傾向が似ているユーザでも，時間的変化が異なる場合には異なるクラスタへクラスタリングされる．提案手法ではユーザの時間的なエントリ行動の変化を考慮したため，就職活動終了時期が遅くなると想定されるユーザの予測精度が向上したと考えられる．

また，本提案の活用事例として予測される就職活動終了時期が最早/最遅のクラスタの特徴分析を行った．最も就職活動終了時期が遅くなると予測されるクラスタに属するユーザは，小売店等の企業に高い嗜好を持ち，3 月までそれらの企業への嗜好が時間が経つにつれて高まっていた．これは早期から業種を絞り，一貫してエントリ行動を変化させないことが就職活動終了時期が遅くなる一要因であると考えられる．一方，最も就職活動終了時期が早くなると予測されるクラスタは，業種が自動車や総合電機で，従業員規模が大きい企業に高い嗜好を持つユーザが所属するクラスタであった．このクラスタのユーザが嗜好を持つ潜在クラスは，理系のユーザに人気である企業が所属するクラスである．また 3 月までにその潜在クラスへの嗜好が薄れていく傾向にあり，就職活動を通じて他の業種や分野の企業へも視野が広がっていることがわかった．以上のように，得られたクラスタを分析することは，就職活動終了時期の予測に対する一助となり，就職活動終了時期に影響を与える要因の把握につながるといえる．

5 まとめと今後の課題

本研究では，就職活動の終了時期が遅くなるユーザを予測するために AM を用いてユーザのエントリ傾向を定量化し，エントリ傾向の時間的変化を考慮したクラスタリング手法を提案した．また，就職ポータルサイトのデータを用いた実験によりユーザのエントリ傾向の時間的変化をモデルに組み込むことの有効性を示した．

今後の課題として，具体的なユーザへのサポートをする手法の提案，クラスタ数の変化に伴う影響の評価などが考えられる．また，就職活動終了時期はユーザの属性に大きく依存していることがわかっているため，これらの情報を考慮したモデルの拡張が望まれる．

参考文献

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