エントリーの時間的順序関係を考慮した 就職ポータルサイトにおける推薦システムに関する研究

経営情報学研究

5212F008-8　大森悠矢
指導教員　後藤正幸

A Study of Recommender System on Internet Portal Sites for Job Hunting Considering Order Relation of Application

OMORI Yuya

1　研究背景・目的

近年，多くの学生が，就職活動を Web 上で支援する就職ポータルサイトを活用している．このような就職ポータルサイトは，サイトを利用する企業から利益を得るビジネスモデルによって成り立っているため，サイト運営企業にとって，個々の企業に対する学生の入社試験申込み (以下，エントリー) 数を確保することが重要となっている．しかし，実際には少数の大企業にエントリーが集中しており，その他多くの企業は相対的にエントリー数が少ないといった現状がある．一方，このような Web サイトには学生のサイト上の行動履歴が蓄積されているため，この大規模データを有効活用することで，被エントリー数の少ない企業に対するエントリー数の向上に向けた施策を行なえる可能性がある．

そこで，本研究では，実際に多くの学生が利用している就職ポータルサイト A を対象事例とし，被エントリー数の少ない企業のエントリー数を向上させることを目的とした企業推薦手法について検討を行う．現状，就職ポータルサイト A では，前年度に就職活動を行っていた学生のエントリー履歴を統計処理し，各学生に対して企業の推薦を行っている．しかし，1 年間のエントリー履歴から計算した企業クラスタ間の相関係数を用いているため，被エントリー数が相対的に多い人気企業であるほど，他企業との相関係数が高くなってしまい，結果として人気企業を推薦してしまう傾向が強くなっている．エントリー数の少ない企業の推薦精度をあげるためには，学生のエントリー嗜好を重視した推薦を行う必要がある．そのためには，学生の局所的なエントリーの順序関係，すなわち「ある企業をエントリーした人は近い将来この企業をエントリーしやすい」といった傾向を考慮し，年間を通じてエントリーされやすい企業を推薦するのではなく，推薦時点から近い将来にエントリーされる可能性の高い企業を推薦すべきである．そこで本研究では，エントリーの局所的な時間的順序関係をモデル化することで，推薦時点における学生のエントリー履歴から，近い将来にエントリーの可能性が高い企業を予測し，各学生に推薦する手法を提案する．従来，時間的な順序関係を考慮した推薦モデルについては，例えば [1]-[4] において提案されているが，いずれも一般的な EC サイトを対象とした推薦モデルである．本研究では，利用する学生が毎年全員入れ替わるといった就職ポータルサイトの特徴を考慮し，前年度の全学生のエントリー履歴からエントリーの時間的順序関係をモデル化し，推薦年度の学生のエントリー履歴に依存した近い将来のエントリー企業を予測，推薦する手法を提案する．

手法の有効性を検証するため，サイト A における過去のエントリーデータを活用したシミュレーション実験と共に，サイト A で就職活動中の学生に対して実際に推薦を行うことによる実証実験を行う．その結果，提案手法により被エントリー数の少ない企業に対するエントリー数を向上させることが可能であることを示す．

2　準備

2.1　エントリーの人気企業への集中

一般的に，毎年，学生のエントリーは一部の企業に集中する傾向がある．図 1 にサイト A における 2011 年 12 月から 2013 年 3 月の各企業の被エントリー数を示す．

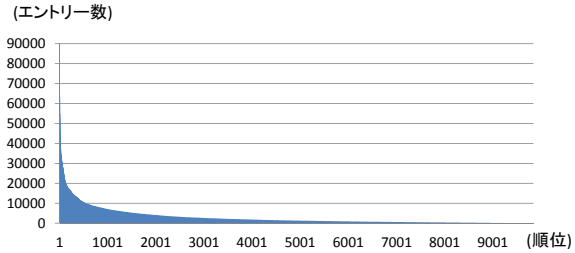

図 1:エントリーの人気企業への偏り

図 1 の横軸は各企業を被エントリー数で降順に並べたときの順位であり，縦軸は各企業の被エントリー数である．最も被エントリー数の多い企業は約 8 万件であり，被エントリー数の多い企業と少ない企業の差が非常に大きく，一部企業にエントリーが集中していることが分かる．被エントリー数の少ない企業にはサイト利用の有用性がほとんどないように感じられてしまうため，就職ポータルサイト運営側にとってこのような偏りがあることは望ましくない．

2.2　エントリーの時間的順序関係

就職ポータルサイト上での学生のエントリー行動には，例えば業種などの「ある特徴を持つ企業をエントリーした直後に別のある特徴を持つ企業をエントリーしやすい」といったエントリーの局所的な時間的順序関係が存在する．また，その順序関係には企業の特徴ごとに差が存在する．以下の表 1 にサイト A で発生している局所的な順序関係の差の具体例を示す．

表 1. エントリーの順序関係の度合い

	政府系統	生活協同組合
政府系統		0.0229
生活協同組合	0.0114	

	建設	建設コンサル
建設		0.062
建設コンサル	0.043	

	食品	農林
食品		0.0299
農林	0.0167	

	医薬品	医療関係
医薬品		0.0412
医療関係	0.0302	

表 1 の各値は，行に示されている業種を持つ企業へのエントリーの後，3ヵ月から 6ヵ月以内に列に示されている業種を持つ企業をエントリーした割合である．なお，エントリー数の多い業種はエントリー割合が必然的に高くなることから，順序関係の程度を定量化するため，各業

種へのエントリー数を業種ごとの総エントリー数で割った値から割合を計算した．また，エントリーデータは学生 17,822 人の 2011 年 12 月から 2013 年 3 月までのサイト A のエントリー履歴 1,048,576 件を使用した．業種数は 126 である．表 1 から見てとれるとおり， 生活協同組合業界をエントリーした後に政府系統業界をエントリーする割合は 0.0114 であるが，政府系統業界をエントリーした後に生活協同組合業界をエントリーする割合は 0.0229 と約 2 倍となっており，0.0115 の差がある．また同様に，建設業界と建設コンサル業界，食品業界と農林業界，医薬品業界と医療関係業界にも順序関係にそれぞれ同程度の差があることが分かる．このように，相対的に平均エントリー割合の低い業種へのエントリー数向上を考えた場合，時間的順序関係を考慮する必要があると考えられる．全体から見た割合は小さくとも，エントリー割合が約 2 倍異なるということは，その業種にとってのインパクトは大きい．

3 従来手法

サイト A の推薦モデルでは，推薦時点の前年 1 年間のエントリー履歴から企業クラスタ間のエントリーの共起関係を推定している．それを用い，各学生に対し直近 10 件のエントリー企業と共起関係の高い企業クラスタを推薦している．企業クラスタとは，業種 $\mathcal{A}$=$\{a_1, a_2, ..., a_I\}$，従業員規模 $\mathcal{B}$=$\{b_1, b_2, ..., b_J\}$，本社所在地 $\mathcal{C}$=$\{c_1, c_2, ..., c_K\}$ の 3 つの組み合わせで構成され，各企業はそれぞれ企業クラスタ $\mathcal{X}$=$\{x_{ijk} | a_i \in \mathcal{A}, b_j \in \mathcal{B}, c_k \in \mathcal{C}\}$ に所属し，推薦は企業クラスタごとに行う．また，企業クラスタごとの共起関係は推薦時点の前年 1 年間のエントリー履歴から業種ごとの相関，従業員規模ごとの相関，本社所在地ごとの相関を計算し，それらの積で求めている．以下に具体的な推薦企業の算出方法を示す．

推薦時点前年にサイト A に登録していた学生を $\mathcal{S}$=$\{s_1, s_2, ..., s_L\}$, 学生 s_l の企業クラスタ x_{ijk} への 1 年間のエントリー数を N_l^{ijk} とする．また，学生 s_l の業種 i の企業クラスタへのエントリー数，学生 s_l の従業員規模 j の企業クラスタへのエントリー数，学生 s_l の本社所在地 k の企業クラスタへのエントリー数を以下のように定義する．

$$N_{i,l}^{\mathcal{A}} = \sum_{j=1}^{J} \sum_{k=1}^{K} N_l^{ijk} \tag{1}$$

$$N_{j,l}^{\mathcal{B}} = \sum_{i=1}^{I} \sum_{k=1}^{K} N_l^{ijk} \tag{2}$$

$$N_{k,l}^{\mathcal{C}} = \sum_{i=1}^{I} \sum_{j=1}^{J} N_l^{ijk} \tag{3}$$

このとき，業種 i の企業クラスタへのエントリー数の平均値 $m_i^{\mathcal{A}}$, 従業員規模 j の企業クラスタへのエントリー数の平均値 $m_j^{\mathcal{B}}$, 本社所在地 k の企業クラスタへのエントリー数の平均値 $m_k^{\mathcal{C}}$ は以下の式で与えられる．

$$m_i^{\mathcal{A}} = \frac{\sum_{l=1}^{L} N_{i,l}^{\mathcal{A}}}{L} \tag{4}$$

$$m_j^{\mathcal{B}} = \frac{\sum_{l=1}^{L} N_{j,l}^{\mathcal{B}}}{L} \tag{5}$$

$$m_k^{\mathcal{C}} = \frac{\sum_{l=1}^{L} N_{k,l}^{\mathcal{C}}}{L} \tag{6}$$

以上のもと，業種 $i, i' \in \mathcal{A}$ のエントリーの相関係数 $R_{i,i'}^{\mathcal{A}}$, 従業員規模 $j, j' \in \mathcal{B}$ のエントリーの相関係数 $R_{j,j'}^{\mathcal{B}}$, 本社所在地 $k, k' \in \mathcal{C}$ のエントリーの相関係数 $R_{k,k'}^{\mathcal{C}}$ は以下の式で求められる．

$$R_{i,i'}^{\mathcal{A}} = \frac{\sum_{l=1}^{L} \left(N_{i,l}^{\mathcal{A}} - m_i^{\mathcal{A}}\right)\left(N_{i',l}^{\mathcal{A}} - m_{i'}^{\mathcal{A}}\right)}{\sqrt{\sum_{l=1}^{L} \left(N_{i,l}^{\mathcal{A}} - m_i^{\mathcal{A}}\right)^2} \sqrt{\sum_{l=1}^{L} \left(N_{i',l}^{\mathcal{A}} - m_{i'}^{\mathcal{A}}\right)^2}} \tag{7}$$

$$R_{j,j'}^{\mathcal{B}} = \frac{\sum_{l=1}^{L} \left(N_{j,l}^{\mathcal{B}} - m_j^{\mathcal{B}}\right)\left(N_{j',l}^{\mathcal{B}} - m_{j'}^{\mathcal{B}}\right)}{\sqrt{\sum_{l=1}^{L} \left(N_{j,l}^{\mathcal{B}} - m_j^{\mathcal{B}}\right)^2} \sqrt{\sum_{l=1}^{L} \left(N_{j',l}^{\mathcal{B}} - m_{j'}^{\mathcal{B}}\right)^2}} \tag{8}$$

$$R_{k,k'}^{\mathcal{C}} = \frac{\sum_{l=1}^{L} \left(N_{k,l}^{\mathcal{C}} - m_k^{\mathcal{C}}\right)\left(N_{k',l}^{\mathcal{C}} - m_{k'}^{\mathcal{C}}\right)}{\sqrt{\sum_{l=1}^{L} \left(N_{k,l}^{\mathcal{C}} - m_k^{\mathcal{C}}\right)^2} \sqrt{\sum_{l=1}^{L} \left(N_{k',l}^{\mathcal{C}} - m_{k'}^{\mathcal{C}}\right)^2}} \tag{9}$$

これらを用い，企業クラスタ x_{ijk} と企業クラスタ $x_{i'j'k'}$ の共起関係 $CO(x_{ijk}, x_{i'j'k'})$ を，以下の式で求める．

$$CO\left(x_{ijk}, x_{i'j'k'}\right) = R_{i,i'}^{\mathcal{A}} \cdot R_{j,j'}^{\mathcal{B}} \cdot R_{k,k'}^{\mathcal{C}} \tag{10}$$

各学生への推薦企業は，各学生の推薦時点直近 10 件のエントリー履歴から代表となる企業クラスタを決定し，その企業クラスタと共起関係の高い企業が対象となる．代表となる企業クラスタは，各学生の推薦時点直近 10 件のエントリー企業のうち，最もエントリーの多い業種，従業員規模，本社所在地を組み合わせた企業クラスタとなる．

4 提案手法

本研究では，被エントリー数の少ない企業のエントリー数向上のため，エントリーの局所的な時間的順序関係を考慮した推薦手法を提案する．推薦により被エントリー数の少ない企業へのエントリー数を向上させるには，推薦時点の学生の嗜好から，近い将来のエントリー企業を正確に予測し，推薦する必要があると考えられる．例えば，学生が被エントリー数の多い企業に対してエントリー行動を行う場合,「個別企業の情報を入念に調査検討のうえ，自身の嗜好と合致するかを見定める」というよりは，人気企業であるという理由でエントリーを行う傾向が強まる．一方，被エントリー数の少ない企業に対してのエントリー行動は，企業の概要や特徴を考慮し，現時点での自らの嗜好に合った企業をエントリーする傾向が高い．そのため，被エントリー数の少ない企業に対して推薦を行う場合，推薦時点における学生の嗜好，つまり局所的なエントリー傾向を適切に表現したモデル化を行うことが必要である．そこで，提案手法ではエントリーの局所的な時間的順序関係を考慮し，近い将来エントリーするであろう企業を予測するモデルを構築する．

相関係数で 1 年間全体の傾向をモデル化する従来手法の場合，被エントリー数の多い企業が推薦されやすくなってしまう．そのため，学生個々の嗜好に合致した多様な企業が推薦されるというよりは，被エントリー数の少ない企業群の中でも相対的に被エントリー数の多い企業が多くの学生に対して画一的に推薦されることになる．すなわち，従来手法は学生個々の嗜好を適切に表現したモデルであるとは言えない．そこで，提案手法では局所的な順序関係を学習することで，各時点における局所的な

エントリー傾向を考慮する．具体的には，提案手法では，学生の直近 D 件のエントリー履歴に対し，エントリーの時系列データを一定間隔で分割したものから算出した企業クラスタ同士の順序関係の高い企業を推薦することで，推薦時点から近い将来にエントリーされる可能性の高い企業を推薦している．提案手法は以下のステップに沿って行われる．

Step1) 各学生のエントリーの時系列データを $2D$ 件間隔で分割する．

Step2) 全学生に対する長さ $2D$ の分割データの集合を $\mathcal{E}$=$\{e_1, e_2, ..., e_G\}$，e_g における前半 D 件のうち企業クラスタ $x_{i'j'k'}$ が含まれる回数を $h^{g,\alpha}_{i'j'k'}$，後半 D 件のうち企業クラスタ x_{ijk} が含まれる回数を $h^{g,\beta}_{ijk}$ とし，e_g における前半 D 件のうち企業クラスタ $x_{i'j'k'}$ が含まれる割合，後半 D 件のうち企業クラスタ x_{ijk} が含まれる割合を

$$p^{g,\alpha}_{i'j'k'} = \frac{h^{g,\alpha}_{i'j'k'}}{D} \tag{11}$$

$$p^{g,\beta}_{ijk} = \frac{h^{g,\beta}_{ijk}}{D} \tag{12}$$

で与え，e_g における企業クラスタ $x_{i'j'k'}$ と企業クラスタ x_{ijk} の順序関係 $\lambda_g\left(x_{ijk}|x_{i'j'k'}\right)$ を以下の式で推定する．

$$\hat{\lambda}_g\left(x_{ijk}|x_{i'j'k'}\right) = p^{g,\alpha}_{i'j'k'} \times p^{g,\beta}_{ijk} \tag{13}$$

これを用い，企業クラスタ x_{ijk} と企業クラスタ $x_{i'j'k'}$ の順序関係を以下の式で推定する．

$$\hat{P}\left(x_{ijk}|x_{i'j'k'}\right) = \sum_{g=1}^{G} \hat{\lambda}_g(x_{ijk}|x_{i'j'k'}) \tag{14}$$

Step3) Step2 で求めた条件付き確率と，推薦時点 t における学生 s_l の直近 D 件のエントリー企業クラスタから，推薦時点 t における学生 s_l の直後 D 件を対象とした企業クラスタ x_{ijk} へのエントリー確率を以下の式で求める．

$$P\left(x_{ijk}|s_l, t\right) = \sum_{i'=1}^{I}\sum_{j'=1}^{J}\sum_{k'=1}^{K} P(x_{i'j'k'}|s_l, t) \cdot \hat{P}(x_{ijk}|x_{i'j'k'}) \tag{15}$$

$P(x_{ijk}|s_l, t)$ はある推薦時点 t から近い将来における，学生 s_l の企業クラスタ x_{ijk} に対するエントリー確率を表している．$P(x_{i'j'k'}|s_l, t)$ はある時点 t における学生 s_l の推薦時点直近 10 件のエントリー企業の割合，$P(x_{ijk}|x_{i'j'k'})$ は企業クラスタ $x_{i'j'k'}$ の後に企業クラスタ x_{ijk} をエントリーする確率，すなわちエントリーの順序関係を表しており，Step4 により前年のエントリーデータから推定される．

Step4) (15) 式で求めた条件付き確率が高い企業クラスタ V 件を学生 s_l に推薦する．

5 実験

提案手法の有効性を示すため，2 種類の実験を行った．1 点目は，サイト A の実データを用いたシミュレーション実験 (実験 1)，2 点目はサイト A で実際に学生に推薦を行うことで従来手法と推薦精度を比較する A/B テスト (実験 2) である．また，サイト A による過去の経験に基づく知見と検証実験により，D は 5 から 10 程度であることが望ましく，それらの差はほとんどないとされている．そこで，従来手法では D=10, 提案手法では計算時間削減のため D=5 として実験を行った．

5.1 実験 1

以下では，サイト A における 2 年分の実データを使用したシミュレーション実験を行い，その結果から提案手法の有効性を示す．

5.1.1 実験条件

実験は，サイト A の 2011 年 12 月から 2013 年 3 月のエントリーデータ (2013 年度採用活動データ) で学習を行い，2013 年 4 月から 2013 年 9 月までのエントリーデータ (2014 年度採用活動データ) の一部でテストを行った．学習データは 2013 年度，テストデータは 2014 年度の採用活動に関するエントリーデータであるので，業種等の企業クラスタは同じであるが，学生ユーザは全て入れかわっている．学習データ，テストデータ共に，業種数 I=126, 従業員規模数 J=8, 本社所在地数 K=49, 企業数 10,304 社, 学習データは学生数 L=559,225, エントリーデータ 31,422,431 件であり，テストデータは学生数 14,999 人のエントリーデータ 648,500 件である．本研究では，エントリー数の比較的少ない企業を対象としているため，テストデータからエントリー数上位企業 W 件のデータを除いて評価することとする．実験は，テストデータに含まれる各学生のエントリーの時系列データを 10 件間隔で分割し，その前半 5 件をモデルへ入力し，推薦された企業を後半 5 件と照合することで推薦の精度を評価する．テストデータのエントリーの時系列を 10 件間隔で分割した時の最後の端数が 1 件から 5 件の場合，照合ができないので，その端数部分はテストデータから取り除いている．実験は推薦企業数 V を 1,5,10,30,50,70,80 と変化させ，W を 5,000 とした場合と，V を 80 に固定，W を 2,000 から 5,000 まで 1,000 単位で変化させた場合の 2 種類行った．

5.1.2 評価指標

本研究は，エントリー数の多くない企業へのエントリー予測精度向上を目的としており，手法を評価するため評価指標として検出率を用いた．検出率とは学生のエントリー企業を推薦できた確率を示す指標であり，以下の式で与えられる．

$$\text{検出率} = \frac{S_V}{Z} \tag{16}$$

S_V は推薦された企業のうち，テストデータの後半 5 件中に存在したものの数であり V によって変化する．Z はテストデータを 10 件に分けた際，後半 5 件にあてはまるエントリーの総数である．

5.1.3 結果・考察

W=5000 に対し，V を変化させた時の検出率を図 2 に示す．

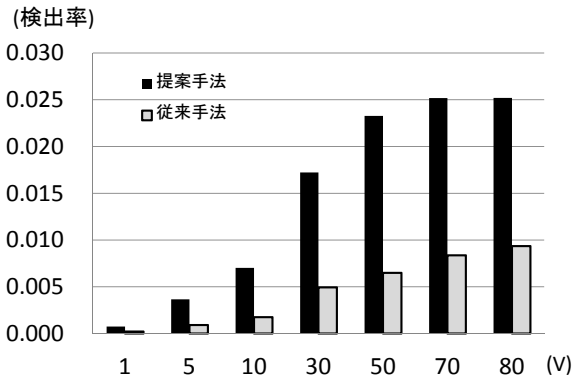

図 2. 推薦企業数 V を変化させたときの検出率

V=80 に対し,W 変化させた時の検出率を図 3 に示す．

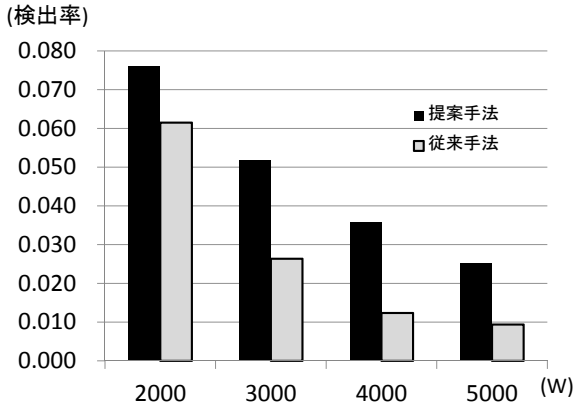

図 3. 除く上位企業数 W を変化させたときの検出率

これより，W を 5000 に固定した場合の全ての V，V を 80 に固定した場合の全ての W に対して提案手法の検出率の方が高いことが分かる．従来手法では局所的な順序関係を考慮していないことから，多くの学生に同じような企業が推薦される可能性があり，また過去や遠い未来にエントリーの可能性が高い企業も推薦されてしまう．一方，提案手法では，従来手法の問題を解決するような局所的な順序関係を考慮した学習を行ったことが，提案手法の検出率が向上した要因だと考えられる．

5.2 実験 2

提案手法で実際に学生に推薦を行った場合のエントリーへの影響を比べるため，サイト A にアクセスした学生をランダムに 2 クラスに分割し，一方には従来手法で推薦，もう一方に提案手法で推薦を行う A/B テストを実施した．

5.2.1 実験条件

学習データは実験 1 と同様であり，推薦を行った期間は 2013 年 9 月 20 日から 10 月 8 日である．また，実験 1 と同様，エントリー数の多くない企業を対象としているため，エントリー数上位 W 企業へのエントリーも除くこととする．実験 2 は，実験 1 において提案手法が特に有効であった W を 5,000 の場合で行った．

5.2.2 評価指標

推薦による効果を正確に評価するため，評価指標は実際にエントリーが行われた数を推薦を見た数で割った値であるエントリー率を用いた．エントリー率は以下の式で与えられる．

$$\text{エントリー率} = \frac{Q}{U} \tag{17}$$

Q はエントリー数，U はユーザが推薦を見た回数を表す．

5.2.3 結果・考察

従来手法と提案手法のエントリー率を以下に示す．

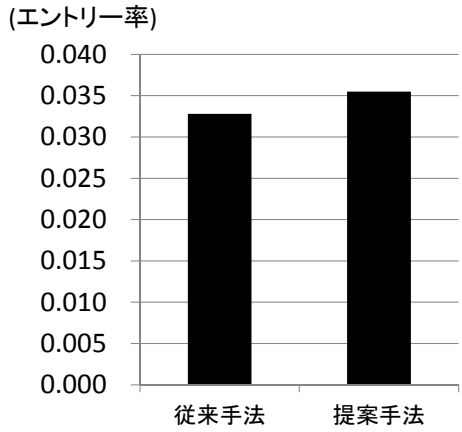

図 4. 実装実験エントリー率
(上位 5000 企業以外へのエントリー率)

図 4 より，若干ではあるが，エントリー率で提案手法が従来手法を上回っていることがわかる．従来手法では 1 年間を通してエントリーされやすい企業を推薦しているため，推薦時点では興味のない企業を推薦している可能性があるが，提案手法では近い将来エントリーされやすい企業を推薦しているため，個々の学生に適した推薦を行えたことが結果に表れたと考えられる．しかし，実験を実施できた期間が 9 月 20 日から 10 月 8 日までであり，比較的遅い時期であると言える．そのような時期にまだ就職活動中である学生は推薦に対してアクティブではない可能性が高く，改めて就職活動の早い時期に実験を行うことが望まれる．

6 まとめと今後の課題

本研究では，エントリーの局所的な順序関係を考慮し，被エントリー数の少ない企業のエントリー数向上を目的とした推薦システムを提案した．評価実験の結果，被エントリー数の少ない企業に対し，エントリー数向上の要因となる推薦時点から近い将来のエントリー企業予測精度を向上できることが明らかになった．また，実際の就職ポータルサイトで実験を行い，被エントリー数の少ない企業に対して，エントリー率が向上することを示した．今後の課題は，エントリー数の多い企業に対して有効な推薦システムの提案や，そのシステムと本研究を組み合わせた推薦システムの構築があげられる．

参考文献

[1] Ding, Y. and Li, X., “Time Weight Collaborative Filtering,” *14th ACM International Conference on Information and Knowledge Management,* pp.485–492, 2005.

[2] Pavlov, D. and Pennock, D., “A Maximum Entropy Approach to Collaborative Filtering in Dynamic, Sparse, High - Dimensional Domains,” *Advances in Neural Information Processing,* vol. 15, pp.1441–1448, 2002.

[3] Song, X., Lin, Y. C. ,and Sun, T. M., “Personalized Recommendation Driven by Information Flow,” *ACM SIGIR,* pp.509–516, 2006.

[4] 川前徳章，坂野鋭, 山田武士，上田修功, “ユーザの嗜好の時系列性と先行性に着目した協調フィルタリング,” 電子情報通信学会論文誌, Vol.J92-D, No.6, pp.767–776, 2009.