論文・事例研究

Ensemble LDAを用いた既存および新規顧客へのスタイリスト推薦

高　正妍，田澤　浩二，チョウ　イ，大原　靖之，
山野上　勇人，桑原　惇，片山　翔太，中田　和秀

1. はじめに

近年，美容院の店舗数が年々増加しており，オーバーストア化している傾向にある [1]．その中で顧客争奪の戦いに勝つためには，顧客が美容院に求めていることを把握することが大切である．リビングくらし HOW 研究所 [2] によると，顧客が美容院を利用するうえで重視するポイントとして「料金」「カットの技術」に次ぎ「スタイリストとの相性」が挙げられている．この顧客とスタイリスト間の相性を定量的に評価できれば，マーケティングをより円滑に行えるだろう．

本研究では，ヘアサロンチェーン店の POS データから顧客とスタイリスト間の相性値の算出モデルを提案する．相性値の算出には潜在的意味解析のモデルである Latent Dirichlet Allocation (LDA) から，より安定した結果を出せるように改善した Ensemble LDA を提案し，これを利用した．LDA を用いることで，単純な判別モデルでは表現できない顧客とスタイリストの潜在的な嗜好性や性格が潜在トピックとして抽出可能になると考えられる．その後，顧客とスタイリストの潜在トピックの類似度を計算することで相性のよさを評価する．また，算出した相性値から指名替えする既存顧客へのスタイリスト推薦を行い，さらに，Factorization Machines (FM) と組み合わせることで新規顧客へのスタイリスト推薦も行う．

提案手法の検証は，経営科学系研究部会連合協議会主催，平成 29 年度データ解析コンペティションで提供されたデータを使用した．

こう　しょうけん，たざわ　こうじ，ちょう　い，おおはら　やすゆき，やまのうえ　ゆうと，くわばら　しゅん，かたやま　しょうた，なかた　かずひで
東京工業大学工学院経営工学系
〒 152–8552　東京都目黒区大岡山 2–12–1
受付 18.7.25　採択 18.11.2

2. 相性値の数値化と Ensemble LDA

提供されたデータの分析から，スタイリストの異動に伴い，顧客の指名状況が大きく変動することがわかった．特に，スタイリストが離職したと考えられる際に，そのスタイリストを指名していた顧客も離反することが多い．実際，あるスタイリストを継続的に指名していた顧客の 70.5%が，そのスタイリストの離職から 3 カ月以内で離脱している．スタイリストの突発的な離職などが起きた場合で，代わりのスタイリストを適切に推薦できれば，それに伴う顧客の離脱を未然に防ぐことができる可能性があり，経営上メリットは大きい．

そのため，本研究では，既存顧客の指名替え先予測を行う．事前分析により，このタスクを顧客属性を用いた判別モデルで予測することは困難であることを確認した．そこで，顧客とスタイリストの潜在的な特性を潜在要因分析手法を用いて分析する．だが，データのスパース性が極めて高く，潜在要因分析でよく使われている PLSA，NMF，LDA はいずれも高い精度が得られなかった．しかし，LDA は Dirichlet 分布を使うことから，PLSA と NMF と違って，学習データに含まれないデータも確率計算ができる．本研究では，このことを利用して学習データのスパース性を削減できることに着目し，LDA を用いた顧客とスタイリスト間の相性値の数値化手法を提案する．

また，美容院に行く約 50%の新規顧客はスタイリストを指名しておらず，リピート率が低い．このため，指名替えする既存顧客のみならず，指名したことがない新規顧客に対しても，彼らに相性のよいスタイリストを推薦することができれば，顧客の満足度とリピート率が上がり，店の売上の向上が期待できる．3.2 節で，LDA を用いた相性値数値化手法の発展として，新規顧客の指名先予測手法を説明する．

本節では，まず，LDA を用いた顧客とスタイリスト間の相性値の数値化方法を提案する．次に，推薦の安

定性向上のための提案モデル Ensemble LDA について説明する．

2.1 LDA

顧客とスタイリストの潜在的な関係を，自然言語処理の技法の一つである潜在的意味解析を用いて表現する．潜在的意味解析のためによく利用されているモデルとして，トピックモデルがある．特に Latent Dirichlet Allocation (LDA) [3] は，一つの文書には複数の潜在トピックが存在すると仮定し，そのトピックの分布を離散分布としてモデル化する統計的潜在変数モデルである [4]．トピック数を K，文書数を M としたとき，Dirichlet 分布のパラメータ $\boldsymbol{\alpha} \in \mathbb{R}^K$, $\boldsymbol{\beta} \in \mathbb{R}^M$ を用いて，トピック分布 $\boldsymbol{\theta}_d = (\theta_{d,1}, \ldots, \theta_{d,K})$ と単語出現分布 $\boldsymbol{\phi}_k = (\phi_{k,1}, \ldots, \phi_{k,V})$ はそれぞれ Dirichlet 分布 $Dir(\boldsymbol{\alpha})$, $Dir(\boldsymbol{\beta})$ に従って生成されると仮定する．そして，変分ベイズ法などの手法で学習した $\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ を用いて，文書 d におけるトピック k の出現確率 $\theta_{d,k}$ とトピック k における単語 v の出現確率 $\phi_{k,v}$ を推定する．$\theta_{d,k} = p(k|d)$，$\phi_{k,v} = p(v|k)$ と解釈すると，文書 d における単語 v の出現確率は以下のように計算できる．

$$p(v|d) = \sum_{k=1}^{K} p(v|k)p(k|d) \tag{1}$$

この LDA を使って顧客とスタイリストの関係について分析を行う場合，顧客を文書 d，スタイリストを単語 v と対応させる．潜在的意味解析を行うモデルの中で，LDA を使う利点として以下の二つが挙げられる．

- 確率分布が Dirichlet 分布に従って生成されているため，学習に含まれていないデータに対し出現確率 $\theta_{d,k}$ を計算できる．一方，PLSA のような学習データから確率を定義している手法，NMF のような学習データ分析のみを行う手法では新規データの分析ができない [3]．
- LDA のオンライン学習アルゴリズムを使う場合，毎回の更新は一部のデータのみを使って行うという特徴を利用し，追加されたデータを用いて既存のモデルを追加的に更新することができる [5]．

2.2 相性値の定義

スタイリスト v のトピック k の出現確率は，ベイズの定理を用いて $p(k|v) = p(v|k)p(k)/p(v)$ と計算することができる．ここで，トピックの出現確率 $p(k)$ は推定された Dirichlet 分布のパラメータ $\boldsymbol{\alpha} \in \mathbb{R}^K$ を用いて $\alpha_k / \sum_{k=1}^{K} \alpha_k$ として計算し，スタイリストの出現確率 $p(v)$ は学習データにおいての各スタイリストの指名された回数の割合である．

このスタイリストのトピック出現確率 $p(k|v)$ と顧客のトピック出現確率 $p(k|d)$ を使い，顧客 d とスタイリスト v の相性値 $A(d,v)$ を次のように定義する．

$$A(d,v) = \sum_{k=1}^{K} p(k|d)p(k|v) \tag{2}$$

この相性値は $0 \leq A(d,v) \leq 1$ の範囲をとる．式 (2) は，顧客 d とスタイリスト v のトピックの構成割合をベクトルにしたとき，それらの内積と一致する．つまり，協調フィルタリングで用いられる二つの特徴ベクトルの類似度の計算に相当している．

2.3 Ensemble LDA

一つの文書に大量の単語がある状況と異なり，1 人の顧客が指名するスタイリストの数はそう多くない．このような極めてスパースなデータを扱うとき，LDA の学習が収束せず，安定した推定結果が得られないことが多い [6]．また，潜在トピック数の増加に伴い，学習データがモデルのパラメータ数と比較して非常に少ない場合，過学習の問題があると指摘されている [7]．そのような状況下で，中村ら [7] は複数の LDA モデルの出現確率 $p(v|d)$（式 (1)）の推定結果の平均を取る Multiple LDA (M-LDA) を提案し，単一の LDA と比較して同程度のモデル規模（＝潜在トピック数 × モデル数）で高精度かつ安定した性能が実現できることを示している．

本研究は，相性値 $A(d,v)$（式 (2)）を計算する LDA モデルに集団学習の枠組みを取り入れた Ensemble LDA (E-LDA) を提案する（図 1）．複数個の単独に学習した LDA の出力から計算された相性値を統合することにより，スパース性の高い学習データからでも安定した推薦結果が得られるようになる．顧客 d とスタイリスト v に対して，$n \in \{1, \ldots, N\}$ 番目の LDA モデルの相性値 $A_n(d,v)$ を式 (2) のように算出し，N 個の相性値の平均値，

$$\tilde{A}(d,v) = \frac{1}{N} \sum_{n=1}^{N} A_n(d,v) \tag{3}$$

を E-LDA の最終的な相性値の出力とする．統合する際にトピックの情報を使っていないことから，それぞれの LDA モデルに違うトピック数を設定して学習してもよいが，本研究では各 LDA のトピック数が同じ場合を扱う．

統合する LDA モデルの数を 1 から 500 まで変化させたとき，各モデル数に対して 15 回の予備実験を行っ

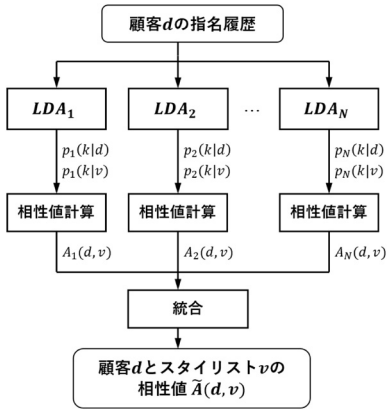

図 1　Ensemble LDA モデルによる相性値計算の流れ

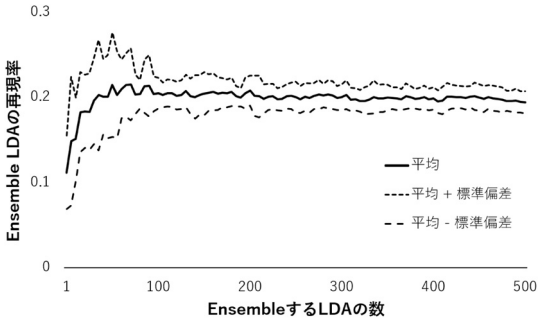

図 2　Ensemble の効果

た．そのときの各モデル数に対する再現率の平均と平均 ± 標準偏差の推移は図 2 のようになる．統合する LDA のモデル数が少ないとき，データの高いスパース性によって，LDA の推定結果が不安定であり，再現率の標準偏差が大きい．モデル数を 250 以上にしたとき，再現率の標準偏差が小さくなり，安定した予測結果が得られるようになっているのが確認できる．

3.　スタイリスト推薦

本節では，まず，E-LDA を用いた既存顧客へのスタイリスト推薦について述べる．次に，E-LDA と FM を組み合わせた新規顧客へのスタイリスト推薦手法を提案する．

3.1　既存顧客へのスタイリスト推薦

すべての顧客もしくは一部の顧客を対象として，各スタイリストに対する指名回数をコーパスに用いた E-LDA を学習する．すると，スタイリストのトピック出現確率と，その指名履歴を学習済みの E-LDA に入力することによりトピック出現確率を計算することができる．一部の顧客を対象にして，各スタイリストに対する指名回数をコーパスに使った場合（4 節の前処理 1），コーパスに含まれない顧客も存在するが，その顧客の指名履歴を学習済みの E-LDA に入力することによりトピック出現確率を計算することができる．その後，式 (3) によって，顧客とスタイリスト間の相性値を計算する．指名替えする既存顧客へのスタイリスト推薦にあたっては，推薦可能なスタイリスト（顧客が通った店舗に所属しているスタイリストの中で，1 年半の学習期間内で合計指名回数が 50 回以上であるアシスタントでないと思われる人）の中で，相性値が一番高いスタイリストを推薦する．

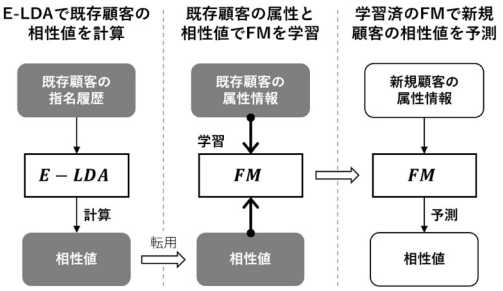

図 3　E-LDA と FM による新規顧客の相性値予測スキーム

3.2　新規顧客へのスタイリスト推薦

既存顧客と違って，新規顧客には指名履歴がないため，上記の方法で顧客のトピック出現確率を計算することはできない．しかし，近年では，美容院をオンラインで予約するウェブサービスを利用する顧客が増えつつあり，ウェブサービスから予約した顧客の性別や誕生年代の属性情報を取得することができる．本節では属性情報を用いた新規顧客への相性値の予測手法を説明する．

新規顧客の相性値予測モデルの全体イメージは図 3 のようになる．まず，既存顧客の指名履歴で学習した E-LDA を用いて，既存顧客とスタイリスト間の相性値を計算する．次に，既存顧客の属性情報と相性値の関係を Factorization Machines (FM) [8] を用いて学習する．予測の際には，新規顧客の属性情報を入力し，学習済みの FM で相性値を計算する．

One-hot 表現した質的データと量的データの両方を同時に扱う属性情報ベクトル $\boldsymbol{x}$，相性値 y を学習の入力とする．今回の推薦タスクでは，属性ベクトル $\boldsymbol{x}$ として，性別，都道府県，店舗までの所要時間などを用いた．入力データは表 1 のようになる．FM をスタイリストごとに学習し，その際にそのスタイリストに対応する相性値を y として用いる．

推定パラメータ $w_0 \in \mathbb{R}$, $\boldsymbol{w} \in \mathbb{R}^m$ および $\boldsymbol{u}_i \in \mathbb{R}^k$ を行としてもつ行列 $\boldsymbol{U} \in \mathbb{R}^{m \times k}$ を用いて，次数 2 の

表 1　FM の入力データ例

	女	男	A 県	B 県	…	時間	相性値
$\boldsymbol{x}^{(1)} =$	(1	0	0	1	…	33)	$y^{(1)} = 0.2$
$\boldsymbol{x}^{(2)} =$	(0	1	1	0	…	40)	$y^{(2)} = 0.7$
$\boldsymbol{x}^{(3)} =$	(1	0	1	0	…	37)	$y^{(3)} = 0.3$

FM の数式モデルは

$$\hat{y}(\boldsymbol{x}) := w_0 + \sum_{i=1}^{m} w_i x_i + \sum_{i=1}^{m} \sum_{j=i+1}^{m} \langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle x_i x_j \quad (4)$$

と定義される．ここで，w_0 は全体のバイアス，w_i は i 番目の属性の重要度を表す．$\hat{w}_{i,j} := \langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle$ は i 番目と j 番目の属性の交互作用を表す．

FM の交互作用パラメータは因子分解可能であるため，データのスパース性が高くてもパラメータはうまく推定できる [8]．この特徴を利用して，学習データにない属性情報の組合せをもつ新規顧客に対しても，属性の交互作用を考慮した相性値予測を行うことができる．今回取り扱うデータは，顧客の総数に対し，偏りが大きく，かつ属性情報の組合せ数が多いという性質をもつ．たとえば，顧客の属性として都道府県を考えると，東京の美容院に対して北海道のデータ例が少なく，ほかの属性との組合せのパターンも極めて少ない．FM を使うことによって，都道府県が北海道になっているデータ例があれば，北海道に関連するバイアスと相互作用のパラメータを推定できる．

FM によって新規顧客とスタイリスト間の相性値を予測した後は，既存顧客へのスタイリスト推薦と同様の手順でスタイリストを推薦する．

4. 美容院の指名履歴データに対する前処理

実データの分析において，前処理は大変重要となることが多い．最初に生データを丹念に調べ，人間の知識・経験などをもとに「よい」データに作り上げておくことが実用的な分析結果につながる．本節では，効果のあった四つの前処理について説明をする．

前処理 1：顧客の選別

美容院では決まったスタイリストしか指名しない顧客は多く，今回使用したデータでも約 90%の顧客が 1 人のスタイリストしか指名したことがない．このような顧客のデータは独立性が高くスタイリスト間の相対比較ができないため，LDA の学習に不適切である．よって，複数のスタイリストを指名したことがある顧客の履歴のみを学習コーパスとして用いる．この処理により，コーパスのスパース性も多少軽減できると考えられる．なお，前述したとおり，コーパスに含まれない顧客に対しても，学習済み LDA を適用することにより顧客のトピック出現確率を計算することができる．

前処理 2：スタイリストの異動の考慮

データを分析した結果，スタイリストの異動についていく顧客が 30.2%，特定の店舗に愛着のある顧客が 23.2%いることが判明した．この美容院の顧客に関する独特な特徴をモデルの学習に取り入れるために，同じスタイリストでも別の店舗で仕事をした場合，LDA の学習時には別のスタイリストとしてコーパスを作成する．あるスタイリスト v が店舗 $s \in S$ に在籍したことがあるとき，そのスタイリストに店舗 s の情報を加え v_s と表す．学習終了後，各店舗においてのスタイリスト v_s の出現確率を統合し，最終的に，スタイリスト v のトピック k の出現確率 $p(k|v)$ を以下のように計算する．

$$p(k|v) = \frac{\sum_{s \in S} p(v_s|k) p(k)}{\sum_{s \in S} p(v_s)} \quad (5)$$

前処理 3：スタイリスト指向の顧客に対する強調

スタイリストが異動したとき，通う店舗を変えてまで同じスタイリストを指名するスタイリスト指向の顧客が存在する．そのような顧客と指名されたスタイリストの間の相性は比較的によいと考えられる．この相性を強調させるために，スタイリスト指向の顧客の指名回数に $\alpha(> 1)$ 倍で補正をかけてコーパスを作成する．本論文の実験においては $\alpha = 3$ とした．

前処理 4：tf-idf による特徴の強調

tf-idf は，自然言語処理分野でよく用いられる文書中に含まれる単語の重要度を評価する手法であり，tf（単語の出現頻度）と idf（逆文書頻度）の積で計算される．顧客の指名履歴に tf-idf を適用した場合，tf は顧客の指名履歴においてのある特定のスタイリストの出現頻度，idf はあるスタイリストが全顧客にどれくらい指名されたかを表す尺度と考えられる．tf-idf 補正によって，多くの顧客が指名するスタイリストの重要度が低くなり，また，ある特定の顧客に頻繁に指名されるスタイリストの重要度が高くなり，独特な相性関係が強調される．

5. 実験結果

本節では 3 種類の実験結果を述べる．まず，提案手法の E-LDA の有用性を検証するため，指名替えする既存顧客へのスタイリスト推薦タスクに，検証期間においての予測精度を示すことで，ほかの手法と比較する．次に，新規顧客へのスタイリスト推薦タスクに，検

証期間においての新規顧客の指名先を予測することで，E-LDA と FM を組み合わせた手法の有効性を確認する．最後に，美容院の指名履歴データに対する前処理の効果を検証する．

5.1 実験データ

本研究では平成 29 年度データ解析コンペティションで貸与されたヘアサロンチェーン店の指名履歴および顧客情報を用いて分析を行う．以下の実験すべてにおいて，学習期間は 2015 年 7 月 1 日から 2016 年 12 月 31 日の 1 年半，検証期間は 2017 年 1 月 1 日から 2017 年 6 月 30 日の半年とする．

5.2 既存顧客へのスタイリスト推薦結果

3.1 節で述べた既存顧客へのスタイリストの推薦タスクの実験を行う．このタスクでは，顧客の指名履歴から各顧客の趣向にマッチしたスタイリストを推薦するということを行いたい．推薦対象顧客は学習期間においてスタイリストの指名を行い，かつ検証期間に学習期間とは異なるスタイリストの指名を連続して 2 回以上行った顧客である．これは指名替えを行い，指名し続けることが，「相性がよい」ことに起因すると想定したためである．評価指標として，

$$\text{指名替えの再現率} = \frac{\text{指名替えを正しく予測した回数}}{\text{対象指名替え顧客数}} \quad (6)$$

を用いる．LDA のトピック数は全店舗数の 12 より小さい値 8，E-LDA のモデル数は 500 を用いた．

比較手法として，以下で説明する NMF，M-LDA，ランダム推薦，人気順推薦の計四つを用いる．

NMF

NMF はデータ行列を非負値制約のもとで，潜在特徴を表す二つの低ランクの行列の積で近似する手法である．文書に適用したとき，潜在特徴をトピックと解釈できる [9]．NMF を予測タスクに適応するとき，行列分解後に得られる各顧客と各スタイリストの特徴ベクトルに対し，それらの内積が一番大きなスタイリストを推薦する．ただし，NMF ではコーパスに含まれない顧客に対し特徴ベクトルは計算できないため，実験時に前処理 1 は適用していない．

M-LDA

M-LDA では，顧客に対するスタイリストの出現確率を式 (1) で計算し，E-LDA と同じモデル数で出現確率の平均を取ることで統合を行い，統合後の出現確率が最大となるスタイリストを推薦する．

ランダム推薦

ランダム推薦は等確率でスタイリストを選択する．

表 2 既存顧客に対する推薦タスクの各手法の再現率と被推薦スタイリスト数

手法	再現率	被推薦スタイリスト数（75 人中）
NMF	0.175	24 人
M-LDA	0.241	25 人
E-LDA	**0.362**	39 人
ランダム推薦（期待値）	0.182	—
人気順推薦	0.126	11 人
実際	—	45 人

人気順推薦

人気順推薦は各店舗での指名回数が一番多いスタイリストを推薦する．

なおすべての手法において，推薦対象となるスタイリストは，顧客が通った店舗に所属しているスタイリストの中でアシスタントでないと思われる人である．各手法での再現率と被推薦スタイリスト数を表 2 に示す．

すべての手法の中で，提案手法である E-LDA の再現率が一番高いことが確認できる．ここで一つ注目すべきことは，人気順推薦がランダム推薦にも劣ることである．一般的な商品推薦タスクにおいて，人気商品を推薦すると比較的よい再現率が得られることが多い．しかし，今回対象としているスタイリスト推薦タスクでは，1 人のスタイリストが対応できる顧客数に制限があり，ある程度指名が分散しているからだと思われる．E-LDA を M-LDA と比較すると，式 (1) を用いた M-LDA で予測を行う場合，すべてのトピックにおいての出現確率 $p(v|k)$ の高い（低い）スタイリストはどの顧客に対しても出現確率が高く（低く）なる．このため，予測結果は指名回数に強く影響され，一部のスタイリストに集中する．それに対して，提案手法で用いた式 (2) では，$p(k|v)$ $(k \in \{1, \ldots, K\})$ はスタイリスト v が得意とする客層構成と解釈できるため，指名回数に依存せずにスタイリストの特性を表せる．このことが E-LDA の再現率の向上につながると考えられる．実際，推薦されたスタイリストの人数をみることにより，NMF や M-LDA では推薦スタイリストが偏っているのに対して，提案手法である E-LDA では実情に即した幅広い推薦を行っていることが確認できる．この点は，一部のスタイリストだけでは顧客を捌ききれないという現実問題に適応しており，提案手法の強みである．推薦対象に制限がある状況では，本研究で提案した「相性値」のような指標を重視して推薦する必要がある．

図 4 に NMF，M-LDA，E-LDA による各スタイリス

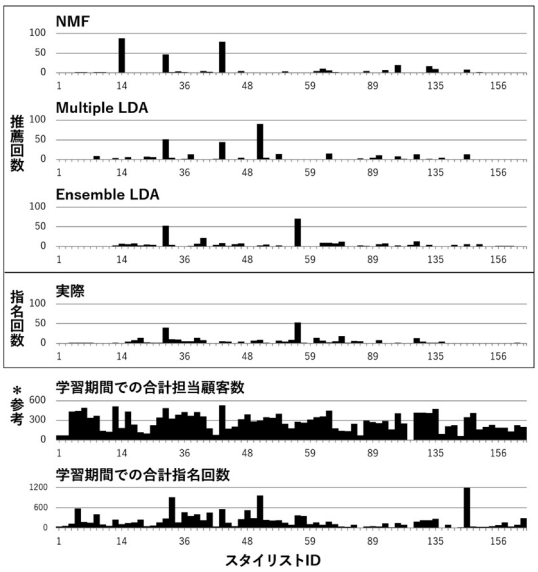

図 4　既存顧客への推薦タスクにおける各スタイリストの被推薦回数

トの被推薦状況を示した．横軸はスタイリストの ID，縦軸は各スタイリストの被推薦回数（または実際に指名替えされた回数）となっている．E-LDA と M-LDA では前処理 2 を適用しているが，図では複数店舗で働いたことのある同じスタイリストを統合した後の結果を示している．参考として，各スタイリストに対する担当顧客数とコーパスに含まれている顧客からの指名回数も載せた．

図 4 より，NMF と M-LDA による推薦状況はスタイリストの人気度（担当顧客数と指名回数の多さ）に影響されていることが読み取れる．具体的には，NMF と M-LDA がともに多く推薦している 43 番のスタイリストは，担当顧客数が 1 番多かった．M-LDA が 1 番多く推薦している 50 番のスタイリストは，学習コーパスに含んでいる顧客の指名総数が 2 番目に多かった．しかし，両者とも実際の指名替えされた回数がさほど高くないことが，NMF と M-LDA の再現率を下げている要因になっている．一方，E-LDA による各スタイリストの被推薦状況は，実際の指名状況に比較的に近い．担当顧客数と指名回数が高くないスタイリストもある程度推薦できている．その中でも特に，57 番スタイリストへの指名替えをうまく予測できていることが読み取れる．

5.3　新規顧客へのスタイリスト推薦結果

3.2 節で述べた新規顧客へのスタイリストの推薦についての実験を行う．推薦対象顧客は，検証期間に初めて指名し，さらに連続して同じスタイリストを指名した顧客を用いる．推薦対象となるスタイリストは，学習期間と検証期間の両方とも勤務履歴をもつ者を用いる．FM で使う属性情報の選別を学習データ上に行い，最もよい組合せを総当たりで探索した．用いた属性は性別，生年代，都道府県，初回来店店舗，DM 送信可否，初回来店店舗までの所要時間，最多来店店舗までの所要時間である．パラメータは E-LDA のモデル数 500，トピック数 8，FM のランク数 10，反復回数 10 回とした．比較として，属性情報と指名回数で学習した FM，ランダム推薦，人気順推薦の 3 種類を用いた．評価指標として，

$$\text{新規指名の再現率} = \frac{\text{新規指名を正しく予測した回数}}{\text{対象新規顧客数}} \quad (7)$$

を用いる．結果は表 3 に示す． 表 3 より，E-LDA による相性値と FM を組み合わせることにより，相性値を使わない FM よりも精度が向上していることがわかる．これは相性値を使うことの有効性を示している．本実験では，顧客属性のスパース性を考慮し，属性から相性値を予測するモデルとして FM を用いた．ただし，ほかの回帰モデルとの比較実験は行っておらず，今後の課題である．

表 3　新規顧客に対する推薦タスクの各手法の再現率

手法	再現率
FM	0.094
E-LDA + FM	**0.206**
ランダム推薦（期待値）	0.164
人気順推薦	0.117

5.4　前処理の効果

次に 4 節で提案した四つの前処理の効果について検証する．表 4 には，前処理 1 から 4 の適用の有無を変えた 16 通りのデータに対し，既存顧客へのスタイリスト推薦タスクを実行したときの再現率を載せている．

コーパスに含める顧客を制限する前処理 1 の効果が大きいことが確認できる．各前処理の効果は互いに影響し合うため単純に比較はできないが，おおむね前処理 1，前処理 2，前処理 3，前処理 4 の順に効果があることが見て取れる．四つの前処理をすべて導入することにより，再現率が 0.159 から 0.362 へと向上しており，前処理を行うことの重要性が確認できる．

6.　おわりに

本研究ではヘアサロンチェーン店の POS データから顧客とスタイリスト間の相性値を算出する Ensemble

表 4　前処理の有無と再現率の関係

前処理				再現率
1	2	3	4	
×	×	×	×	0.159
×	×	×	○	0.124
×	×	○	×	0.171
×	×	○	○	0.121
×	○	×	×	0.197
×	○	×	○	0.141
×	○	○	×	0.171
×	○	○	○	0.153
○	×	×	×	0.203
○	×	×	○	0.253
○	×	○	×	0.235
○	×	○	○	0.224
○	○	×	×	0.250
○	○	×	○	0.274
○	○	○	×	0.253
○	○	○	○	**0.362**

LDA を提案した．さらに，これを用いて既存顧客に対するスタイリスト推薦モデルの提案を行った．相性値に基づいた推薦を行うことにより，推薦対象の供給に制限がある状況で実用的な推薦を行うことが可能となった．また，集団学習の枠組みを取り入れることで，データの疎性に起因する不安定さを解消することに成功した．実務で生じる商品推薦タスクにおいて，商品供給に制限がある，あるいは商品と顧客の関係を表すデータに疎性があることは多い．そのような場合に，本手法は特に有効であると考えられる．なお，[10] では顧客と商品との間の相性値が既知であるとして，同じ商品を推薦する個数を限定した全顧客への推薦商品最適化問題に取り組んでおり，本研究と関係が深い．

提案モデルの実用性をより上げるため，次の三つが考えられる．

1. 提案手法は推薦を人気のスタイリストに集中させないことに特徴があるが，この問題はまだ完全に解消できていない可能性がある．スタイリストの担当制限を直接的に考慮したほうがより再現率の高い推薦ができる可能性がある．
2. データの制限によって，今回は推薦システムにスタイリストの属性情報を利用することができなかった．もしスタイリストの年齢などの情報があれば，それらを提案手法に組み込むことで，再現率を上げることができる可能性がある．
3. 今回の新規顧客へのスタイリストタスクにおいて，属性情報として「都道府県」を用いた．しかし，関東地方の店舗に対する分析で，極めてデータ数の少ない地方のデータは特異的な結果となる恐れがある．そのため，「東北」や「関西」で地方のデータをある範囲でまとめたほうが，より正しく推薦できる可能性がある．

これらについては今後の課題としたい．

指名履歴のない新規顧客へのスタイリスト推薦では，顧客の属性から相性値を推測する Ensemble LDA と FM を組み合わせたモデルを提案した．回帰モデルと組み合わせることにより，相性値の利用範囲を広げることに成功している．

提案した相性値はスタイリスト推薦以外にも利用できる可能性がある．たとえば，スタイリストの店舗再配置を行う際，顧客にとって相性のよいスタイリストがいないという状況を回避できる．

参考文献

[1] 厚生労働省,「美容業概要」, http://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/kenkou/seikatsu-eisei/seikatsu-eisei03/06.html（2018 年 6 月 30 日閲覧）

[2] リビングくらし HOW 研究所,「女性（2014 年/全国）『美容院についてのアンケート』」, https://www.kurashihow.co.jp/wp-content/uploads/2014/07/02cf75b6bef347228a6ee977c5fe46c9.pdf（2018 年 6 月 30 日閲覧）

[3] D. Blei, A. Ng and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, **3**, pp. 993–1022, 2003.

[4] 佐藤一誠,『トピックモデルによる統計的潜在意味解析』, コロナ社, 2015.

[5] M. Hoffman, D. Blei and F. Bach, "Online learning for Latent Dirichlet Allocation," *Advances in Neural Information Processing Systems*, pp. 856–864, 2010.

[6] 坂本俊輔, "超スパースなデータに対する潜在クラスモデルを用いた推薦システムに関する研究," http://www.it.mgmt.waseda.ac.jp/results/student1/2013-M2-Sakamoto.pdf（2018 年 6 月 13 日閲覧）

[7] 中村明, 速水悟, 津田裕亮, 松本忠博, 池田尚志, "複数モデルの統合による LDA トピックモデルの高精度化とテキスト入力支援への応用," 情報処理学会論文誌, **50**, pp. 1375–1389, 2009.

[8] S. Rendle, "Factorization machines," In *Proceedings of 2010 IEEE International Conference on Data Mining*, pp. 995–1000, 2010.

[9] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, **401**, pp. 788–791, 1999.

[10] 土谷拓人, 西村直樹, 鮭川矩義, 高野祐一, 中田和秀, 松本健, "大規模な推薦商品最適化問題に対する確率的劣勾配法," 日本オペレーションズ・リサーチ学会春季研究発表会アブストラクト集, pp. 194–195, 2014.