事例研究［論文］

行動特性に基づいた育児支援サイトユーザの類別化による仮想的フォロー機能の提案

鈴木　優耀，渡邉　晃大，村上　竜之介，齊藤　史哲

1. はじめに

男女共同参画が求められる現代社会においても家庭内における家事・育児の負担の割合は依然として母親の方が多くなりがちであることが指摘されている [1]．昨今における Web サービスの発展に伴い，ネットを通じた女性の出産育児にまつわるストレス・不安の軽減や適切な情報提供は今後ますます重要視されると考えられる．こうした中で，出産育児に関するストレス・不安の低減を目指して Twitter や Facebook などの SNS 上での情報の流通や出産や育児に関する情報サイトなどを通じた情報提供が近年盛んに行われている．本研究の解析対象であるコネヒト株式会社が運営するママリ [2] は情報共有や交流を通じて育児を支援する最も代表的な育児支援サイトの一つである．

一般的な SNS ではユーザ間のフォロー機能や履歴に基づいた情報推薦機能を用いることでユーザ間でのやり取りや情報共有が活発に行われている．その一方で，出産育児に関するユーザの悩みには身体的・生理的な内容を含むセンシティブな投稿が多いことから，ママリではユーザ間のトラブルを回避するためにフォロー機能はあえて設けていないとのことである [2]．ユーザの安心感を維持するうえでこのような設計思想は，医療に代表されるプライバシーに踏み込んだ悩みに関する投稿を扱う同様なサービスにおいても，今後広く受け入れられると考えられる．

フォロー機能を設けないことでユーザ間トラブルや不満を減らすことができる一方で，従来の SNS でなされてきた積極的な情報共有は限定的なものになる可能性もある．ユーザ間での情報共有における真価を発揮するうえでは，情報を届けたい相手や情報を受け取りたい相手を設定できるフォロー機能に準ずる情報推薦機能は重要である．閲覧履歴や投稿数に基づいた一般的な情報推薦のみではユーザ間の価値観や興味対象の近さ，話題の共通性，求められる回答の質といった情報発信者・回答者の相性や素養も考慮した情報提供を行うことができていなかった．本研究は，ユーザ間の直接的な接触を回避しつつ質問や回答内容の情報共有を活発化させる枠組みの構築を目指し，データに基づいたユーザの仮想的フォロー機能を提案するものである．

ここでは，投稿ジャンルや検索数などの行動特性データに基づいてユーザの特徴量を算出し，クラスタリングを通じて各ユーザの特性に適したユーザの投稿内容を提示する枠組みを新たに提案する．提案法により，回答者側のユーザがもちあわせている情報と質問者側のユーザがもちあわせている情報のマッチングを行うことで，フォロワーとフォロイーの関係（フォローする人とされる人の関係）に起因する対人トラブルを回避しつつ，円滑な情報提供の実現が期待できる．たとえば，検索結果や未回答なレビューを表示するときに当該の仮想フォロイーユーザの投稿を上位の投稿として目立つ場所に提示されやすくするなどの方法で適切な接点の提示が期待できる．

先行研究に目を向けると，フォロイー推薦に関する研究はこれまで盛んになされてきているが，その多くが SNS を対象としたものであるがゆえに，既にフォローしているユーザ間の関係性に基づいて新たなフォロイーを推薦するものや，投稿内容に基づいて興味対象が近いユーザを推薦するものが一般的である [3–6]．本研究の対象サービスにはそもそもフォロー機能が存在しないことから従来のフォロイー推薦では活用できていた情報はもちあわせていないが，対象が Q&A サイトであることから，情報を求めて質問をするユーザのニーズと情報を提供したいユーザのニーズは一定数存

すずき　まさあき
千葉工業大学大学院先進工学研究科
〒 275–0016 千葉県習志野市津田沼 2–17–1
わたなべ　こうだい，むらかみ　りゅうのすけ，さいとう　ふみあき
千葉工業大学先進工学部
〒 275–0016 千葉県習志野市津田沼 2–17–1
受付 22.7.15　採択 22.11.4

在する．その情報を活用することで有用なフォロイー推薦に近い情報提供を実現できると考えられる．これにより，情報を求めて質問する傾向が強いユーザに対して回答を通じた情報提供を好むユーザをマッチングすることでコミュニケーションや情報共有の活発化が期待できる．

2. 解析データについて

本研究では，経営科学系研究部会連合協議会主催の令和 3 年度データ解析コンペティションにてコネヒト株式会社より提供された情報サイト「ママリ」におけるデータを用いている．提供されたデータは，妊活・妊娠・出産・子育ての疑問や悩みを解決する情報サイトに関するデータであり，メインユーザである新生児・乳幼児の母親からの投稿検索データやその履歴といった多様な情報が含まれている．

本研究において解析対象としたデータは，提供されたデータの内 2020 年 1 月 1 日から 2020 年 12 月 31 日までの期間における質問データと，それに対する回答データならびに，その期間内における検索履歴データである．本研究では，ユーザの産前・産後・子育てなどといったフェーズの変化に伴う利用内容の変動範囲をある程度抑えることと，処理環境の都合上アクティブユーザを一定数に抑えることを目的に対象データの期間を 1 年に固定している．ここでは，積極的に利用しないユーザや新規ユーザ，休眠ユーザのデータは十分な情報が含まれていないと判断し，これらを除くために期間内における総検索回数が 100 回を超えたアクティブユーザを対象として解析を行っている．また，データ内には膨大な種類のクエリが存在することから，対象とした検索クエリは検索総数上位 500 クエリを対象としている．

3. 特徴量について

3.1 特徴量の構成

本研究はユーザの行動特性に基づいたクラスタリングを通じてユーザを類別化するものであり，各ユーザに対する特徴量を求める必要がある．ここでいう行動特性とは，Q&A サイト利用時におけるユーザがサイト内で取りうる行動の中で実際に取った行動を表している．ユーザは何らかの情報を欲するときは他者に対して質問をするか検索をすることで情報収集を行う．逆に，情報を他者に向けて発信したいときは質問に対して回答することや回答するべき質問を検索する．さらに，「グッドアンサー」の付与を通じて他者回答を評価することで情報を提供する．このようなユーザが取りうる行動を定量化することで，「情報を発信したいユーザなのか？」「情報を収集したいユーザなのか？」また，「どのジャンルの情報を欲しているのか？」「どのジャンルについて詳しいのか？」「回答情報の質の良し悪し」といったユーザの特徴を定量的に表現できる．

提供された質問データには予め用意された質問ジャンル 18 種類の内，いずれか一つのラベルが付与されている．その内，本研究で解析対象とするアクティブユーザ群において実際に使用されていたジャンルは「妊娠・出産」「子育て・グッズ」「サプリ・健康」「ココロ・悩み」「妊活」「家事・料理」「お金・保険」「雑談・つぶやき」「お仕事」「ファッション・コスメ」「家族・旦那」「お出かけ」「産婦人科・小児科」「住まい」「その他の疑問」の 15 種類であった．条件を満たすユーザ群内で利用されていなかったジャンルは「教育・習い事」「避妊など」「出産報告」であった．以上を踏まえて i をユーザ，j を対象のジャンルを識別するためのインデクスとしたとき，あるユーザにおけるジャンルごとの質問数を $q_{ij} \in \mathbb{N}$，回答数を $a_{ij} \in \mathbb{N}$ と表現すると，15 ジャンル存在することからユーザごとの質問，回答に関する行動のベクトルはそれぞれ

$$\boldsymbol{q}_i = (q_{i1}, q_{i2}, \ldots, q_{i15})^T \tag{1}$$

$$\boldsymbol{a}_i = (a_{i1}, a_{i2}, \ldots, a_{i15})^T \tag{2}$$

と表現できる．さらに，回答したコメントに対して獲得されたグッドアンサー数を $g_{ij} \in \mathbb{N}$ とすると，同様に

$$\boldsymbol{g}_i = (g_{i1}, g_{i2}, \ldots, g_{i15})^T \tag{3}$$

として表せる．

また，検索クエリの種類を区別するインデクスを t としたとき，ユーザ i のクエリごとの検索数を $e_{it} \in \mathbb{N}$ と表すと，これはユーザごとの検索クエリを対象とした頻度行列，すなわち Bag of search queries であり，語彙数が次元となる高次元疎ベクトルである．検索総数上位 500 件の検索クエリを対象としていることから，ユーザごとの検索行動ベクトルは次式のような計算をすることで $s_{it} \in \mathbb{R}$ と表すことができる．

$$\boldsymbol{s}_i = (s_{i1}, s_{i2}, \ldots, s_{i500})^T \tag{4}$$

$$s_{it} = \mathrm{IDF}(t) \left[\frac{e_{it} \cdot (k_1 + 1)}{e_{it} + k_1 \cdot (1 - b + \frac{b|D|}{avgdl})} \right] \tag{5}$$

$$\mathrm{IDF}(t) = \log \frac{N - n(t) + 0.5}{n(t) + 0.5} \tag{6}$$

s_{it} は，情報検索尺度 BM25 [7] をユーザごとの検索数に適用したものであり，ユーザごとの検索数には大きな個人差が含まれることからクエリ数の偏りを補正しつつ，その特徴を評価している．ここでは，$|D|$ は各ユーザの検索総数，$avgdl$ は平均検索数（本来の情報検索では文書長を用いるが，ここでは総検索数をユーザの文書とみたてて補正に利用している）を表しており，対象データセット内において $n(t)$ は検索クエリ t を利用した経験があるユーザ数を，N は全ユーザ数を表している．また，k_1 および b はパラメータであり，最も広く利用されている $k_1 = 1.2, b = 0.75$ を用いている．

これらのベクトルはいずれも提供データをユーザごとに集計することで獲得できるものであり，これらを用いて行動特性を評価する特徴量を構築する．以下 3.2〜3.4 節にて本研究で利用する行動特性に関する 3 種類の特徴量（情報取得に関する特徴量，情報発信に関する特徴量，興味の対象に関する特徴量）それぞれの計算方法について説明する．

3.2 情報取得に関する特徴量

Q&A サイトでは，疑問の解消や所望の情報を入手する手段として質問投稿と検索が挙げられる．質問の投稿は，検索結果として所望の情報にたどり着けないときに情報をもつ他のユーザに直接質問することで必要な情報を得ることができるばかりでなく，情報にたどり着くまでの検索によるわずらわしいプロセスを回避することもできる．また，質問投稿は回答者である他のユーザとのやり取りを通じたコミュニケーションの手段とみなすこともできる．他者とのコミュニケーションを通じた悩みの共有は，ユーザの性格によっては，出産育児におけるストレス・不安の解消に貢献する要因といえる．

ここでは，ユーザが求めている情報の程度を式 (1) と同じ指標を用いて情報取得に関する特徴量 $\boldsymbol{X}_i^Q$ として扱う．

$$\boldsymbol{X}_i^Q = \boldsymbol{q}_i \tag{7}$$

質問の回数を調べることで，情報に対する姿勢として情報を求めているユーザであるか否か，さらに，他者とのコミュニケーションを求めるタイプであるか否かを評価できる．また，質問対象のジャンルを評価することでそのユーザが情報を求めている興味の対象も併せて評価できる．

3.3 情報発信に関する特徴量

次に，情報発信に関する特徴量の扱いについて述べる．Q&A サイトにおける質問への回答による効用は回答者自身がもちうる情報を他者に提供することで他者の疑問や不安を解消することにある．求められるユーザは情報を他者に提供することを好み積極的に情報を提供する性格であり，かつ，質問内容の回答として適切な情報をもちうるユーザである必要がある．

回答回数の多寡でジャンルごとの対応の大きさを評価することができるが，単純に多く回答を投稿すればよいわけではなく，その回答の質，すなわち，その回答がどの程度他者にとって役に立ったか？ を評価する必要がある．そこで，本研究では次式のとおりジャンル内においてユーザがグッドアンサーを獲得する割合をグッドアンサー率 $r_{ij} \in \mathbb{R}$ として計算することで評価する．ここではベクトルの対応する要素ごとの除算であるアダマール除算の演算子 $\oslash$ を用いてこれを次のように表すことができる．

$$\boldsymbol{r}_i = \boldsymbol{g}_i \oslash \boldsymbol{a}_i \tag{8}$$

ただし，未回答ジャンルが存在する場合は分母が $a_{ij} = 0$ になるケースを無視して $r_{ij} = 0$ としている．

前の計算結果を用いて情報発信に関する特徴量を以下のように定義する．これはグッドアンサー率を用いてジャンル別にユーザ回答の質の高さを数値化し回答数に対する重み付けをしている．これも同様にベクトルの対応する要素ごとの積であるアダマール積の演算子 $\odot$ を用いて次のように表すことができる．

$$\boldsymbol{X}_i^A = \boldsymbol{a}_i \odot (\boldsymbol{r}_i + \boldsymbol{\epsilon}) \tag{9}$$

これは，ジャンルごとの回答数に対してグッドアンサーの取得率で重み付けをしたものであり，$\boldsymbol{\epsilon}$ の各要素は回答数が多いにもかかわらずグッドアンサーを取得できないユーザに対する評価値が零値になることを避けるための小さな値である．

3.4 興味対象に関する特徴量

情報の取得や発信に関する情報はジャンルのみであり，ユーザの詳細な興味対象が表現できているとは言い難い．そこで，上記の特徴量に加えて，検索履歴の行動特性を用いてユーザの興味を特徴量 $\boldsymbol{X}_i^S$ として表現する．式 (4)〜(6) で計算された BM25 を要素とする $\boldsymbol{s}_i$ を行ベクトルとした行列 S に対して非負値行列因子分解 (NMF) [8, 9] を適用することでユーザの興味対象に関する情報を因子（トピック）として抽出する．

NMF は非負制約を設けた行列分解を通じてデータ

行列を基底行列 U と係数行列 V の積に分解することで次元を削減する機械学習手法である．本研究では，特徴量を $\boldsymbol{X}_i^S$ としているので，特徴量行列を X^s としたとき，

$$S = X^s V \tag{10}$$

として表現できる．これは，分解の結果として得られた低次元に縮約された基底行列 U をユーザの興味を表す特徴量行列 X^s として扱うことを意味している．

4. ユーザクラスタリングに基づいた仮想的フォロイーの選出

先述のとおり，本研究ではユーザの仮想的フォロー機能を提案するものである．フォロー対象を選出するにあたり，類似したユーザを特徴に基づいてクラスタリングする必要がある．この結果に基づいて仮想的フォロワーのニーズに合致したユーザのクラスタを仮想的フォロイーとして対応付ける．

3.2〜3.4 節で求めた特徴量を，次式のように情報取得に関する特徴量 $\boldsymbol{X}_i^Q$，情報公開に関する特徴量 $\boldsymbol{X}_i^A$，興味対象に関する特徴量 $\boldsymbol{X}_i^S$ の順にマージすることでユーザの行動特性ベクトル $\boldsymbol{X}_i$ を表現する．

$$\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{X}_i^Q \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{X}_i^A \\ \boldsymbol{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{X}_i^S \end{pmatrix} \tag{11}$$

この特徴量に基づいてユーザをクラスタリングすることで，類似した特性をもつユーザのセグメントを獲得することができる．

Sakurai et al. [10] は，このように複数のベクトルを結合した連想対として自己組織化マップを学習させ，その結果に対して結合前のベクトルを用いて勝者ノードを決定することで連想記憶モデルを構築している．本研究では，この枠組みをユーザクラスタリングに拡張しつつ，情報公開の特徴量と情報取得の特徴量のマッチングに応用することで，推薦対象ユーザの対応付けを行う．近年においてもセントロイドベースのクラスタリング手法は顧客クラスタリングにおいて広く利用されており [11, 12]，クラスタの特徴を代表するセントロイドベクトルの抽出が容易であることから本研究では k-means に基づいたアプローチを提案する．これらの方法の中でも計算時間が少なく，初期値依存性の影響を抑えて安定的な結果が得られることから本研究では k-means++ を採用している．これにより，クラスタリングの結果として得られたセントロイドベクトル，すなわち，クラスタの重心を用いてユーザのニーズとフォロイー候補クラスタの特徴を対応付ける．ここでは，セントロイドベクトルはクラスタ数だけ存在し，次元数は特徴量と同じであることから上記の特徴量と同様に，クラスタを識別するインデクスを c とすると，次式で表すことができる．

$$\boldsymbol{W}_c = \begin{pmatrix} \boldsymbol{W}_c^Q \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{W}_c^A \\ \boldsymbol{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{W}_c^S \end{pmatrix} \tag{12}$$

クラスタリングの結果として得られたセントロイドベクトルを用いて，フォロイー候補とフォロワー候補となるクラスタを選定する．フォロワーユーザが所属するクラスタの添え字を u としたとき，その重みベクトル $\boldsymbol{F}_u^r$ は情報を求めていることから情報取得に関する重みと興味対象の重みを用いて次式によってあらわすことができる．

$$\boldsymbol{F}_u^r = \begin{pmatrix} \boldsymbol{W}_u^Q \\ \boldsymbol{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{W}_u^S \end{pmatrix} \tag{13}$$

同様に，フォロイー候補クラスタの重みベクトルはもちあわせている情報を表現する必要があることから，情報公開に関する重みと興味対象の重みを併せて次式のようにあらわせる．獲得された v 番目のクラスタの中心に対応するセントロイドベクトル F_v^{ee} は次式によって表現できる．

$$\boldsymbol{F}_v^{ee} = \begin{pmatrix} \boldsymbol{W}_v^A \\ \boldsymbol{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{W}_v^S \end{pmatrix} \tag{14}$$

適切な対象クラスタを提示するにあたり，興味対象が近く，なおかつ情報を提供するジャンルと情報を求めるジャンルも近いクラスタの対応付けをこれらのコサイン類似度に基づいて行う．

$$sim(\boldsymbol{F}_u^{ee}, \boldsymbol{F}_v^r) = \frac{\boldsymbol{F}_u^{ee} \boldsymbol{F}_v^r}{\|\boldsymbol{F}_u^{ee}\| \|\boldsymbol{F}_v^r\|} \tag{15}$$

本研究では，上記の指標が閾値 θ を超えたクラスタに所属するユーザをフォロイーの候補として提示するものであり，これは質問に関するニーズと回答に関するニーズのマッチングを行うプロセスである．

5. データ分析

5.1 分析の設定

対象データを用いた解析を行うにあたり，興味対象の特徴量として扱ったトピック数（クエリ行列を縮約した次元数）を 30 として特徴量を計算している．情

表 1　Bag of search queries から抽出されたトピックの特徴の一例

手法	検索クエリ
Topic 1	陣痛，新生児，おしるし，おしゃぶり，魔の 3 週目，お宮参り，新生児 げっぷ 出ない，里帰り
Topic 2	高温期 X 日目（X は 10〜13），排卵検査薬，生理予定日，妊娠 超初期 症状，妊娠検査薬，妊活
Topic 3	離乳食，生後 6 ヶ月，BCG 跡，予防接種，お食い初め，おしゃぶり，抱っこ紐，うつ伏せ寝
Topic 4	性別，つわり，ベビーナブ，つわり いつまで，胎動 いつから，R 先生，つわり いつから，性別 いつ
Topic 5	離婚，旦那，不倫，ママ友，幼稚園，虐待，義母，旦那 イライラ，シングルマザー，貯金
Topic 6	1 歳 X ヶ月（X は 0〜8），突発性発疹，トイトレ いつから，慣らし保育，イヤイヤ期，断乳，卒乳
Topic 7	生後 X ヶ月 離乳食 Y（X は 6〜8），（Y は「量」または　「食べない」），生後 X ヶ月 生活リズム（X は 5〜9）
Topic 8	慣らし保育，保育園，育児休業給付金，育休延長，育休，保育料，育休手当，突発性発疹，保育士

報公開の特徴量を算出するにあたり式 (9) の計算で用いた ϵ の値は 0.1 としている．クラスタリング時の特徴量はすべて標準化することで無単位化しデータのスケールを統一している．k-means++ のクラスタ数 k は 40 としてユーザクラスタリングを行った．獲得されたクラスタからフォロイーユーザを提示するにあたり，コサイン類似度の閾値 θ は 0.7 としている．これらのハイパーパラメータは，事前にいくつかの設定で解析をためしており，その中でも解析を進めるうえで現実的に扱える値を設定している．たとえば，クラスタ数やコサイン類似度の閾値を大きくすればより精緻なマッチングができるが適切なマッチング対象が見つからないクラスタが多数発生するなどの課題が発生する．逆にこれらの値を小さくしすぎると対象範囲のユーザが広くなりすぎてマッチングが荒くなる．NMF の次元数（基底数）においても同様な性質がある．ここでは特に得られた知識（トピック）の内容の解釈しやすさに重点を置いた設定にし，5.2 節にて結果を示している．

また，ハイパーパラメータと提案システムの挙動の関係性を調べるために $k = 10, 20, 40, 60$ のそれぞれの設定において $\theta = 0.5, 0.7$ とした際に対応付けられたクラスタ群の組み合わせ数を確認した．ここでは，それぞれの設定において 10 施行実施した際の結果をまとめている．また，推薦対象として抽出されたクラスタにおけるグッドアンサーの取得率とクラスタ数の関係性を調べるために $\theta = 0.7$ において $k = 10$〜60 を対象とした実験を行い，5.2 節にてこの結果を示している．

5.2　分析結果

Bag of search queries より興味対象の特徴量として獲得されたトピックの一部を表 1 に示している．これは NMF で分解された係数行列の重みが特に大きい検索クエリを表にまとめており，式 (10) の係数行列 V の重みにおいてそれぞれの検索クエリやその関連語に対して強く反応するトピックの情報を表している．Topic 1 は生まれて間もない新生児をもつユーザに，Topic 2 は妊娠前後のユーザに，Topic 3 と Topic 6 は生後半年を過ぎ離乳食を食べ始めるころの赤ちゃんを育てているユーザに，Topic 4 は胎児の性別が気になり産前に知りたいユーザに，Topic 5 は配偶者などの身の回りや生活に対するストレスに関するトピックに，Topic 7 は赤ちゃん育児に関する情報を集めるユーザに，Topic 8 は保育園に通う程度に成長した幼児を育てているユーザにそれぞれ対応している．

図 1，2 はクラスタ間の類似度行列であり，図 1 はクラスタ数 $k = 40$，図 2 はクラスタ数 $k = 10$ とした際の結果を表している．対角成分はフォロイー候補とフォロワー候補が同一クラスタであるためコサイン類似度が大きな値となっている．ただし，比較しているベクトルは情報公開に関する要素と情報取得に関する要素が異なるものであるため，完全に一致しているとは限らず対角要素でも濃淡が生まれている．すなわち，同一クラスタであったとしても質問数が多いジャンルと回答数が多いジャンルが異なるクラスタではマッチングの指標が小さくなることから，そのようなクラスタでは薄い色になっている．縦の並びは各クラスタのフォロワーユーザの特徴量（式 (13)），横の並びは各クラスタのフォロイーユーザの特徴量（式 (14)）を表しており，それぞれクラスタ間のコサイン類似度の行列を表している．また，対角要素以外の成分においても，類似度が高いクラスタをところどころ確認できる．これは，他のクラスタに所属していて異なるタイプのユーザであったとしてもフォロワー候補が他者に聞きたいジャンルの情報と，フォロイー候補がもちあわせているジャンルの情報が一致していることを表している．一般に，クラスタ数 k を増加させることでユーザのセグメント，すなわちクラスタとして抽出されたユーザ群がより細かく区分され同一セグメントに所属するユーザ間の特徴はより類似したものになる．これを踏

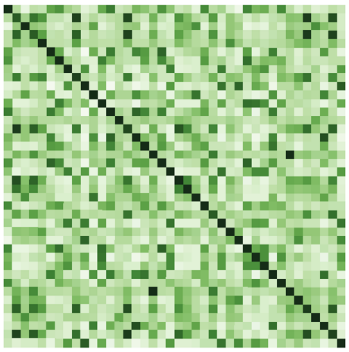

図 1　クラスタ間のマッチングの結果（クラスタ数 $k = 40$ における当該特徴量のコサイン類似度）

図 2　クラスタ間のマッチングの結果（クラスタ数 $k = 10$ における当該特徴量のコサイン類似度）

まえて図 1，2 を比較すると，そのセグメント同士のマッチングのバリエーションの増加が確認できる．これは，クラスタ数が少ないとセグメント化が荒くなり，適切な特徴をもたないユーザがセグメント内に含まれる可能性が高くなることを意味している．逆に，よりセグメントの細分化が進むと適切な推薦候補のセグメントが増えマッチングのバリエーションが増えるといえる．

図 3 はクラスタリングの結果および位置関係を視覚的に把握できるようにするために，データを 2 次元平面上にマッピングすることで可視化した結果を表している．ここでは t-SNE などの広く知られた次元削減法によるデータの二次元配置よりも高速かつセグメント間の配置を明瞭に表示できることから UMAP [13] を採用し，共通するクラスタに属するデータを同一色としてクラスタごとにクラスタ数分の色（計 40 色）に彩色している．セントロイド同士のマッチングにより，色の異なるセグメントの中で，ニーズが一致したクラスタ同士を対応付けたことになる．

表 2 では提案法によって対応付けられた仮想的なフォロワーとフォロイーの関係を表している．ここでは，セントロイドの重みベクトルの値が大きい変数を上から順に並べて表示している．この結果より，類似する質問と類似する回答の対応付けができていることが見て取れる．異なるクラスタ間で，質問のニーズと回答のニーズ（もちあわせている情報）が一致しているものがある一方で，クラスタ 4 のように同一クラスタ内で質問と回答のニーズがきれいに一致しているものも確認できた．ここで際立つジャンルは「妊活」や「仕事」であり，このクラスタは仕事と出産の両立に対する悩みを抱えるユーザ群であるといえる．このようなユーザは同一クラスタ内で似た悩みに対して質問・相談と回答両方行うことで，悩みを共有することで自

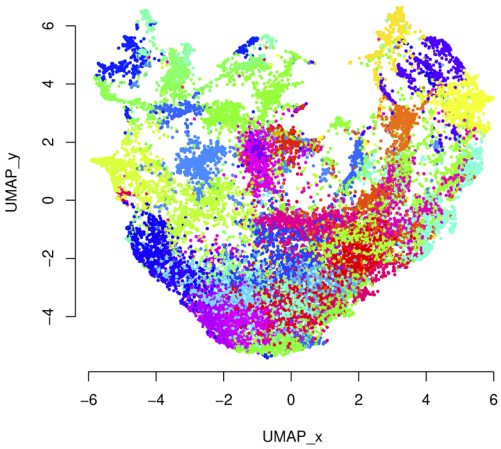

図 3　クラスタリング結果の可視化（クラスタ数 $k = 40$ の結果をクラスタごとに 40 色で彩色）

身の不安やストレスを軽減できる可能性がある．

5.3　挙動の確認

図 4 は提案法でマッチングされたユーザ間でのやりとりにおけるグッドアンサーの取得率を評価している．横軸はクラスタ数 k を，縦軸はグッドアンサーの取得率を表しており，各クラスタ内でのやり取りにおいて取得された分布を箱ひげ図として表示している．ここでは，全体的には k を増加させてユーザのセグメントを細分化することでグッドアンサー取得率が高いセグメントも得られる可能性が示唆された．その一方で，いずれの設定においてもグッドアンサーが全く取得できていないクラスタも存在することが確認でき，このような状況では仮想的フォロイーの推薦結果においてグッドアンサー取得には貢献できないことも確認された．今回対象として選ばれたアクティブユーザにおける全投稿の平均グッドアンサー取得率が 0.51% であったことを考えると，中には 8% を超えるものもあり，ユーザの対応付け次第では一定の効果は認められると

表 2　マッチング結果の例（フォロイーとフォロワーの関係）

フォロワー候補	上位の変数 X_j^Q	フォロイー候補	上位の変数 X_j^A
クラスタ 20	「避妊など（ジャンル 9)」 「住まい（ジャンル 17)」 「ファッション・コスメ（ジャンル 12)」 「その他（ジャンル 99)」	クラスタ 15	「家族・旦那（ジャンル 13)」 「避妊など（ジャンル 9)」 「住まい（ジャンル 17)」 「ファッション・コスメ（ジャンル 12)」
クラスタ 4	「お仕事（ジャンル 11)」 「産婦人科・小児科（ジャンル 15)」 「妊活（ジャンル 5)」	クラスタ 4	「お仕事（ジャンル 11)」 「妊活（ジャンル 5)」 「産婦人科・小児科（ジャンル 15)」
クラスタ 36	「ココロ・悩み（ジャンル 4)」 「教育・習い事（ジャンル 7)」 「家族・旦那（ジャンル 13)」 「サプリ・健康（ジャンル 3)」	クラスタ 27	「教育・習い事（ジャンル 7)」 「サプリ・健康（ジャンル 3)」 「妊活（ジャンル 5)」 「家族・旦那（ジャンル 13)」

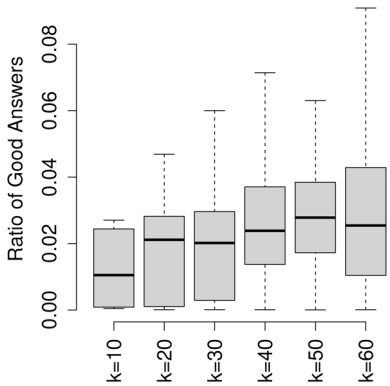

図 4　仮想フォローにおけるクラスタ数とグッドアンサー率の関係性

考えられる．

表 3 は閾値 θ とクラスタ数 k において，推薦対象としてマッチングされたクラスタ（対象ユーザ群）の組み合わせ数を表している．これは，図 1，2 の比較からもわかるとおり，より細分化された顧客群が構築できることからクラスタ数が増加するにつれて推薦対象クラスタの組み合わせが増加することを示している．クラスタ数 k を増やし続けるとそれに伴ってマッチングのバリエーションは余すところなく増加させることができることから仮想フォロー関係におけるユーザ間のニーズの一致度の向上が期待できるが，その一方でセグメント内に所属するデータ数が減少し，推薦される候補者がもちあわせる情報の多様性が失われる可能性も高まる．また，θ は推薦対象のマッチングを決定するための閾値であることから，この値を小さくすると選定基準が甘くなり仮想フォロー対象の組み合わせの数が大きくなる．逆に θ を大きくすると推薦対象の選定基準が厳しくなり仮想フォローの組み合わせの種類が少なくなる．推薦対象としてフォロワーのニーズにより強く当てはまるフォロイーを厳選する際は θ を大きくし，フォロワーのニーズにずれるユーザが含まれる可能性をある程度容認しつつもより広範囲のユーザに対して推薦機能を提供したい際は θ を小さくするなどの施策が必要となる．以上を要するに，実用を想定する際は管理者が閾値 θ やクラスタ数 k を調整することで推薦件数の必要数やその範囲を状況に応じてコントロールする必要があるといえる．

表 3　パラメータ θ, k と推薦対象ユーザ群数の関係性

k		10	20	40	60
$\theta = 0.7$	平均	5.3	19.3	52.4	72.7
	標準偏差	2.0	4.1	8.6	9.8
$\theta = 0.5$	平均	18.9	62.0	165.9	247.2
	標準偏差	4.4	10.4	13.2	10.3

5.4　まとめと考察

実験結果として類別化されたユーザのクラスタを調べると，質問回数が多い割に回答数が極端に少ないユーザや，その逆のパターンのユーザ，すべてにおいて極端に多く対応しているユーザなどさまざまなタイプのユーザがおり，投稿対象のジャンルもユーザによって複数のタイプが確認できた．これらの対応関係や提示先をユーザ間の仮想フォロー機能で提示しあうことができれば，より活発な情報共有の促進が期待できる．また，提案システムは行動履歴に基づいた推薦システムであるが故に，コールドスタート問題，すなわち，新規ユーザや休眠ユーザに対処する際に十分なデータがまだ存在しないため適用できないという問題が常につきまとう．実応用を検討する際は，ユーザに対して興味ジャンルやデモグラフィクス情報を事前に問い合わせ，類似するユーザのセグメントに割り当てた状態から利用を開始するなどの対応が必要と考えられる．

その一方で，実験結果では ϵ の値によって投稿数の影響を抑えたが，回答内容の質よりも量を優先するケースではこの値を大きくすることで対応ができる．今後はこの値の結果に及ぼす影響を精査していく必要があると考えられる．ただし，$\epsilon = 0$ とした回答内容の質を重視する状況では，投稿数の多寡が強く影響してしまうため，スパム投稿などの質の低いものを多く推薦してしまう可能性がある．

また，今回対象としたデータのユーザは出産育児をする母親が対象ということで，子供の成長や産前産後といったステージに応じて求める情報の変遷が激しいと考えられる．今回は 1 年間として期間を区切ったため影響は少ないが，今後はその変化をトレースしながら推薦対象のフォロイーを変更できる枠組みの検討が必要であると考えられる．これに対しては，活動期間を半年程度で区切りながら解析の区間を設定し，同様な解析を繰り返し行うことでユーザの所属クラスタの移り変わりを注意深く確認することでより精緻なユーザの対応付けが実現できると考えている．

6. おわりに

本研究ではユーザの行動特性から特徴量を構築し，ユーザクラスタリングを通じてクラスタにおけるニーズに合致したクラスタ同士をマッチングすることで，ユーザ間仮想的なフォロー構造を提供する方法を提案した．育児支援サイトのようなセンシティブな内容を多く含むサイトでは，このようなアプローチは情報共有活性化の観点において一定の成果が期待できる．提案アプローチの枠組みは産前・産後の女性のストレス軽減の観点において，適切な対応関係を維持したまま間接的なコミュニケーションの促進が期待できる．

今後の課題として，ユーザクラスタリングの粒度，クラスタリング手法の検討を進めていくことが挙げられる．クラスタサイズを増やすことや，一度割り当てられたクラスタ内で再度クラスタリングするアプローチなどが考えられる．また，計算資源が許す状況であれば，Affinity Propagation などのクラスタ数の自動決定を行いつつセントロイドに対応するサンプル (exemplar) を見つけ出すクラスタリング手法の有効性を検討する必要がある．

謝辞　データをご提供いただいたコネヒト株式会社様，データ解析コンペ関係者の皆様，有益なコメントを提供いただいた査読者様に厚く御礼申し上げます．また，本研究は科学研究費（基盤 C）19K04887 より支援いただきました．

参考文献

[1] 内閣府男女共同参画局，『共同参画』，令和 2 年 9 月号，2020.
[2] アンドエンジニア，「「ママリ」は一期一会だからこそ温かい！ コネヒト株式会社 CTO が語る，コミュニティアプリを健全に保つ仕組みと技術」，https://and-engineer.com/articles/YBe2uhAAACYAjM29（2022 年 6 月 1 日閲覧）
[3] 熊本忠彦，灘本明代，“共通話題に対する感情的態度の類似度に基づくフォロイー推薦，”電子情報通信学会論文誌 D, **J100-D**, pp. 500–509, 2017.
[4] 熊本忠彦，鈴木智也，“Twitter ユーザの印象選好を可視化するシステムの設計と評価，”電子情報通信学会論文誌 D, **J98-D**, pp. 788–801, 2015.
[5] H. Chen, X. Cui and H. Jin, “Top-k followee recommendation over microblogging systems by exploiting diverse information sources,” *Future Generation Computer Systems*, **55**, pp. 534–543, 2016.
[6] A. Tommasel, A. Corbellini, D. Godoy and S. Schiaffino, “Personality-aware followee recommendation algorithms: An empirical analysis,” *Engineering Applications of Artificial Intelligence*, **51**, pp. 24–36, 2016.
[7] S. Buettcher, C. L. A. Clarke and G. V. Cormack, *Information Retrieval: Implementing and Evaluating Search Engines*, The MIT Press, 2010
[8] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, **401**, pp. 788–791, 1999
[9] D. D. Lee and H. S. Seung, “Algorithms for Non-negative Matrix Factorization, In *Proceedings of Conference on Neural Information Processing Systems (NIPS)*, **13**, pp. 556–562, 2000
[10] N. Sakurai, M. Hattori and H. Ito, “SOM associative memory for temporal sequences,” In *Proceedings of the 2002 International Joint Conference on Neural Networkss (IJCNN'02)*, pp. 950–955, 2002.
[11] Z. Sun, T. Zuo, D. Liang, X. Ming, Z. Chen and S. Qiu, “GPHC: A heuristic clustering method to customer segmentation,” *Applied Soft Computing*, **111**, 107677, 2021.
[12] Y. Li, X. Chu, D. Tian, J. Feng and W. Mu, “Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm,” *Applied Soft Computing*, **113**, 107924, 2021.
[13] L. McInnes, J. Healy and J. Melville “UMAP: Uniform manifold approximation and projection for dimension reduction,” arXiv, https://arxiv.org/abs/1802.03426v3, 2018.

