DEIM2021 E31-2

生花ECサイトの購買履歴に基づく商品特性分析モデル

山極　綾子† 　雲居　玄道† 　後藤　正幸†

† 早稲田大学　〒 169-8555 東京都新宿区大久保 3 丁目 4 番 1 号
E-mail: †saxophone-0105@ruri.waseda.jp

あらまし　近年，購買履歴データをマーケティング施策に活用する研究が多くなされている．その多くは顧客と商品の共起関係に基づき商品の特徴量を分析するモデルであるが，同一顧客による被購買数が少なく，かつ購入の度に嗜好が変化するような商品群には適用することが難しい．本研究ではこのような特徴を持つ事例である生花 EC サイトを対象とし，購買履歴データに基づく商品特性分析手法を提案する．具体的には，購買履歴データが商品，購入用途，顧客属性の組み合わせで与えられることに着目し，購入用途と顧客属性が商品の購買有無に与える影響を商品特性と見なすことで商品間の類似度を評価する．ここでそれら関係性のモデル化には，説明変数間の交互作用を考慮可能な Factorization Machine を用いる．最後に，提案手法を実データに適用し，その有効性を示す．

キーワード　機械学習，購買データ，マーケティング，Factorization Machine，商品特性，類似度

1　背景と目的

近年，インターネット技術の発達と PC やスマートフォンなどの情報端末の普及により，多くの購買行動がインターネット上の EC サイトで行われるようになった [1]．さらに，それら EC サイト上で行われた購買履歴データは各企業に蓄積されており，その大量のデータをビジネスへ活用する重要性が高まっている [2]．例えばデータを活用したマーケティング施策の例として，顧客一人一人の嗜好に合わせた 1on1 マーケティングが注目されている [3]．しかし，1on1 マーケティングを人手によってのみ導入することは，コストの観点から難しい．そこで，機械学習を活用して各顧客の嗜好を明らかにし，マーケティングに活用する手法が注目されている．

顧客の嗜好に着目したマーケティング施策のための 1 つの観点として，例えば顧客の嗜好に基づき商品の類似性分析を行い，類似した商品を顧客に推薦することによる売り上げ増加を目的とするものがある．類似性の評価方法として，商品のジャンルなどの情報を用いるなどの単純な手法に加え，大量に蓄積されたデータを活用する新しい手法が多く研究されている [4]．後者の手法である Item2Vec [5] は近年，EC サイト上での音楽や映画，ゲーム，日用品など様々な商品群の購買履歴に適用され，その有効性が示されている [6]．この手法は，顧客の嗜好は時系列によって変化することを仮定し，短期間に購入された商品を同じ嗜好に基づき購入された類似商品であると見なし，埋め込み空間上に商品を表現するモデルである．そのため，この手法は同一顧客からの購入数が少ない商品群や，購入の度に嗜好が変化する商品群に対しては適用することができない．

一方，EC サイト上での購買行動は，日用品など頻繁に消費される商品に留まらず，他者への贈答用の商品でも行われるようになっている．贈答用商品を販売する EC サイトの事例として，本研究では生花 EC サイト A 社から提供された購買履歴データを対象とする．A 社が運営する生花 EC サイトを利用することにより，購入者は希望する商品を，希望する受領者に指定した日時に届けてもらうことができる．ここで，生花 EC サイト上での顧客の購買行動は，年に 1 回のみの “母の日”や “誕生日”など特定のイベントでの贈答を用途として購買行動が行われることが多く，その購入間隔は長い場合が多い．そのため A 社は，既存顧客に対し，一度購買した用途に加え別用途でも購買を促すことを重要視しており，実際にそのための商品の提示が行われている．しかし現状は担当者の経験に基づき，各用途の注目商品を一律に提示するに留まっている．加えて，連続した購買であっても，贈答品としての購買では受領者が異なる可能性があり，その場合は購買行動が同一の嗜好に基づいていないことも考えられる．つまり，本研究の対象事例である生花 EC サイトは，購入間隔が長く，かつ購買の度に顧客の嗜好が変化する可能性があるため，従来手法では類似性の推定が難しい．

本研究では，対象事例のような購入間隔が長く，その嗜好が変化しやすい商品群に対して，顧客の嗜好を反映した商品の類似性分析手法を提案する．ここで，同じ商品 A を購入した顧客の嗜好は類似していると仮定し，それ以外の商品を購買した顧客との違いを商品特性と見なし，その値を用いて商品間類似度を評価することを考える．具体的には，商品 A を購買したデータを正例，それ以外の商品を購買したデータを負例，顧客属性や購入用途を説明変数として二値分類器を学習し，推定された説明変数の係数を商品の特性と見なす．ここで，得られた購買履歴データをそのまま用いた場合，正例と負例のデータ数が極端に偏るため，負例データのランダムサンプリングが必要となるが，負例の選択方法によって学習される係数が変化することに注意する必要がある．また，同じ顧客属性であっても用途が変われば受領者が異なり，嗜好が変化する可能性がある．つまり，顧客属性と用途の間には関係性が存在している．そこで，二値分類器として，説明変数間の交互作用を表現すること

が可能なモデルである Factorization Machine（以下，FM）[7] を用いる．最後に，実際のデータに提案手法を適用しその有効性を示す．

2 準　　備

本研究では，生花 EC サイト A 社から提供された購買履歴データを対象とする．本章ではまず A 社のビジネスモデルとデータの概要について説明した後，関連研究について述べる．

2.1 生花 EC サイトビジネスモデル

A 社のビジネスモデルを以下の図 1 に示す．

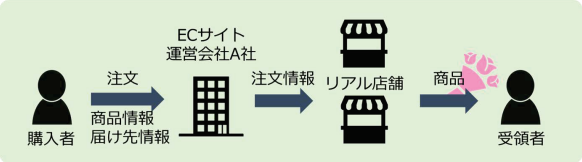

図 1　生花 EC サイト A 社のビジネスモデル

顧客は生花 EC サイト上で購入する商品を選び，その注文情報と共に受領者の情報を入力する．その後，A 社に加盟するリアル店舗のうち，受領者に商品を届けるために最も適切な店舗を A 社が選択し，その店舗に注文情報を送る．最後に，リアル店舗が受領者に商品を届けることで，購買行動が終了する．

この生花 EC サイト上での購買行動の多くは，“母の日”や“誕生日”といった何らかのイベントに際して，購入者が受領者に対し贈答品を送るために行われる．そのため，商品も顧客の購入目的となりうるイベントに基づき制作されており，A 社が定めたカテゴリと紐づけられて管理されている．例えば，“母の日”の贈答用に制作された商品には，カテゴリとして “母の日” が付与されている．また，生花 EC サイト上では，各カテゴリごとに，イベントに合わせた特集ページが作成されている．そのため，顧客がある用途でサイトを訪れ，目的に合致した特集ページから商品を選ぶ場合，選択肢に入る商品のほとんどに，同一のカテゴリが付与されている状況がある．

ここで，生花 EC サイトでの購買行動について，多くの EC サイトと異なる点が 2 つ挙げられる．1 点目として，購買間隔が長いことが挙げられる．生花 EC サイトでの購買行動の多くは，他者に対する贈答品であった．それらのきっかけとなるイベントは，主に年に 1 回のみ発生するものであり，多くの顧客は年に 1 回程度の購買に留まっている．そのため，既存顧客に対し，一度購買した用途に加え別用途でも購買を促すことが重要となる．実際に A 社ではそのための商品の提示が行われているが，現状は担当者の経験に基づき，各用途の注目商品を一律に提示するに留まっている．異なる点の 2 つ目は，購買行動ごとに，商品を選ぶ際の嗜好が変化する可能性である．日用品や消耗品の購買行動について考えた場合，短期間に購買された商品は，顧客が同じ嗜好の元で購入したと考えることが自然である．一方，生花 EC サイトでの購買においては，連続する購買であっても，贈答品としての購買においては受領者が異なる可能性があり，その場合は購買行動が同一の嗜好に基づいていない場合がある．例えば，“母の日”用途で，あるカーネーションを使った商品を購買した顧客 A が，次の機会には友人の誕生日プレゼント用に，全く異なるアレンジメント花束を購入する，といった場合が考えられる．そのような場合には，商品を選択する際の嗜好が変化するといえる．

2.2 事 前 分 析

本研究では，対象購買履歴データが商品，購入用途，顧客属性の組み合わせで与えらえることに着目し，商品特性の推定を行う．ここで，同じ購入用途でも顧客属性が異なれば受領者の属性も変化すると考えられるため，嗜好も変化すると考えられる．すなわち，購入用途と顧客属性の間には関係性が存在するといえる．実際に，購入用途と顧客属性の関係性を図 2 に示す．

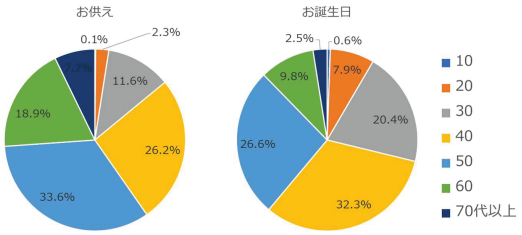

図 2　購買用途と顧客属性（年代）の関係

左の円グラフは購買用途が “お供え”である購買，右は “お誕生日”用途の購買について，それぞれの購買顧客の年代分布を示している．例えば “お供え”用途では最も購買している顧客の年代は 50 代であるが，お誕生日用途では 40 代が最も多くなっており，顧客属性の一つである年代傾向は購買用途によって異なっている．つまり，用途と顧客属性の間には関係性があり，それらを適切に表現することが，商品特性のモデル化に重要であると考えられる．

そのような，購買履歴データを活用し，顧客の嗜好に基づく商品類似度を分析する手法が多く提案されている．

2.3 Factorization Machine

本研究では，購買履歴データに二値分類器を適用し，商品特性の推定を行うことを考える．ここで，購買履歴データに含まれる “顧客属性”，“購入用途”の間には交互作用が存在していた．そのため，本研究で用いる二値分類器として，比較的少ないパラメータ数で入力データ中の説明変数間の交互作用を考慮することができ，高い予測性能を示す予測モデルとして知られている Factorization Machine [7](以下，FM) を用いる．

FM を二値分類器として用いた場合の，定式化について説明する．いま，目的変数を $y \in \{-1, 1\}$，説明変数ベクトルを $\boldsymbol{x} = (x_1, x_2, ..., x_d)$，$(x_i \in \{0, 1\}, i = 1, 2, ..., d)$ としたデータが与えられているとし，w_0 をバイアス項，$\boldsymbol{w} = (w_1, w_2, ..., w_d)$ を各説明変数の重みベクトルとする．特徴量間の交互作用をその積により表現することができる交互作用行列を $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_d)^\top \in \mathbb{R}^{d \times k}$ とし，その要素を交互作用ベクトル $\boldsymbol{v}_i = (v_{i1}, v_{i2}, ..., v_{ik})(k \ll d)$ と定義する．

ここで，説明変数 x_i は 0, 1 のいずれかをとる二値変数である必要がある．例えばアンケートデータのように，各説明変数に対し当てはまる場合は 1，そうではない場合は 0 といった

データであれば，そのまま入力データを用いることができる．一方，本研究の対象事例のように，入力データの各説明変数に対し，例えば “顧客性別”として “女性 (F)”，“男性 (M)”，“その他 (O)”などそれぞれ異なる値を持っている場合には，各説明変数ごとに 1hot ベクトル化し，FM を適用する必要がある．

N 件の学習データに対し，各説明変数の種類ごとに，1hot ベクトルに変換されたものが FM における説明変数ベクトル $\boldsymbol{x}$ となる．また目的変数 y について，本研究では商品 A を購買したデータと，それ以外の商品を購買したデータを識別する分類器の学習を行う．そのため，目的変数とする購入商品についても同様に変形し，y_i^A をある商品 A を購入したとき 1，それ以外の商品を購入した時 -1 をとなる変数とする．

いま，上述で定義した変数を用いると，$\boldsymbol{x}$ が与えられたとき，FM は次の式 (1) で表される．

$$f(\boldsymbol{x}) = w_0 + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle x_i x_j \tag{1}$$

ただし，第 3 項の $\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$ は式 (2) 定義される内積である．

$$\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \sum_{l=1}^{k} v_{il} v_{jl} \tag{2}$$

FM は重回帰モデルに対して式 (1) の第 3 項を加えることで，説明変数間の交互作用を表現している．ここで，FM の交互作用は，$d \times k (k \ll d)$ の交互作用行列と呼ばれる行列の積を計算することで，重回帰モデルに比べ比較的少ないパラメータ数で交互作用を表現することができる．そのため，FM は交互作用が存在する問題に対して重回帰モデルよりも予測精度の向上が期待できる．

3 提 案 手 法

対象とする生花 EC サイト上の購買履歴データの分析手法の着想について述べたあと，提案手法のアルゴリズムについて述べる．

3.1 着　　想

本研究で対象とする生花 EC サイトでは贈答品を主に取り扱っている．ここで，そのきっかけとなるイベントは主に年に 1 回のみ発生するものであり，実際に，年に 1 回のみの購入に留まる顧客が全体の約 7 割を占めている．つまり，購買行動の特徴として同一顧客による購入間隔が長いことが挙げられる．そのため，既存顧客に対し，一度購買した用途に加え別用途でも購買を促すことが重要となる．実際にそのための商品の提示が行われているが，現状は担当者の経験に基づき，各用途の注目商品を一律に提示するに留まっている．ここで，顧客の嗜好に合致した商品を提示することができれば，購買行動を促すことができると考えられる．そこで本研究では，類似した嗜好を持つ顧客に購買されるか否かという視点から，その類似度を評価する手法を提案する．それにより，ある用途で商品 A を購入した顧客が他の用途で購入する可能性の高い商品を提示することが可能となり，顧客個人に合わせたマーケティング施策に結びつけることが可能となる．

対象事例では購入間隔が長く，また購入の度に顧客の嗜好が変化するため，従来研究が行われてきた購買有無に基づく行列分解や潜在クラスモデル，埋め込み表現などの商品類似性評価手法では適切に評価を行うことができない．そのため，購買履歴データが商品，購入用途，顧客属性の組み合わせで与えられることに着目し，ある商品を購入したデータと，他の商品を購入したデータを分類する二値分類器の係数を用いて商品特性を評価することを考える．同じ商品 A を購入した顧客の嗜好は類似しており，かつその嗜好は顧客属性に現れると仮定すると，得られた係数が類似している商品は，その商品を購入する顧客の嗜好に基づき類似していると見なすことができる．実際に，佐和 [8] は,「係数そのものが説明変数の重要度を表すわけではないが，目的変数との関係性を示すものである」と述べている．

ここで，商品 A の購買有無を分類器の目的変数とした場合，対象商品以外のすべての商品を購買したデータを負例として扱うことになり，データ数が極端に偏ってしまう．このような正例と負例の数が偏った学習データに対し二値分類器を適用した場合，適切に係数の推定を行うことができない．なぜならば，二値分類器では学習データの目的変数と予測値が一致するような係数を学習するため，学習データが偏っている際には，多い方の目的変数を出力しやすい係数が学習されてしまうからである．そのようにして得られた係数は，真にデータの特徴を表すものとはいえない．そのため，適切に負例データをサンプリングし，正例と負例の数を一致させる必要がある．また，負例のサンプリングについて，選択されたデータによって得られる係数が異なることに注意する必要がある．例えば，ある商品 A と B がそれぞれ異なる用途で主に購入される場合，これら 2 つの商品を分類するには用途が何であるかが重要であり，購入用途のみを用いても二値分類器の学習が可能となるため，学習される係数には顧客の嗜好が反映されないと考えられる．つまり，適切な負例の選択を行うことで，二値分類器の係数に顧客の嗜好を反映することが可能であり，結果として係数を用いた商品特性の評価が可能になると考えられる．

また，本研究では二値分類器の説明変数として購入用途と顧客属性を用いるが，それらの間には交互作用が存在していた．仮に，すべての説明変数間の組み合わせについて個別にその交互作用を学習しようとする場合，学習が必要なパラメータ数が非常に多くなってしまい，学習が適切に行われない可能性がある．そこで本研究では，交互作用行列の積により交互作用を表現することで，少ないパラメータで交互作用を評価することが可能な FM [7] を二値分類器として用いることとする．

3.2 提案アルゴリズム

本研究では，顧客の嗜好に合致した商品を提示するマーケティング施策の一助とするため，その時の購入用途に合わせ，過去に顧客が購入した商品と類似した商品を提示することを可能とする，顧客の嗜好を反映した商品の類似性分析手法を提案する．ここで，対象事例における購買行動について，購買間隔が長いこと，また顧客の嗜好が変化しやすいことが特徴として

挙げられた．そこで，同じ商品 A を購入した顧客の嗜好は類似していると仮定し，商品 A の購買履歴データと，それ以外の商品の購買履歴データを識別する，二値分類器の導入を考える．具体的には，説明変数に “顧客属性” と “購入用途” を用いることで，各商品との関係性を係数の形で推定することができ，さらに係数を特徴量と見なすことで，商品の類似度を定義することができる．また，“顧客属性” と “購入用途” の間には関係性が存在するため，本研究では二値分類器として，少ないパラメータで交互作用を表現することが可能な FM を用いる．

ここで，二値分類器は正例と負例のデータを分類する係数を学習するため，負例の選択方法によって得られる係数は変化する．すなわち，顧客の嗜好を反映した係数を学習するためには，どのデータを負例として用いるかが重要であるということである．そこで，顧客が商品 A を購入した際に検討対象としたと考えられる商品を負例に選択することで，顧客の嗜好を表現することを考える．具体的には，同じカテゴリの商品が同ページに提示されているという EC サイトの設計を活用し，顧客が対象商品 A の比較対象とするであろう商品を，各商品のカテゴリに基づき選択することで，適切な負例を選択することを考える．最後に，分類器で得られた係数を用いて商品類似度を算出する．

次に，提案手法のアルゴリズムを以下に示す．

（1） 学習用データセットの作成

（2） 分類器の学習

（3） 係数を用いた類似度評価

まず，正例とした商品の購入時に，顧客が比較対象としたであろうと考えられる商品群を負例対象商品とし，ランダムサンプリングを行うことで，学習データセットを作成する．次に，FM を用いて係数の推定を行う．なお，本研究では FM の学習手法として交互最小二乗法 [9] を用いている．最後に，求めた係数を用いてコサイン類似度を算出し，得られた商品類似性について，高い類似性を持つ商品がどのような特徴を持っているか分析を行う．ここで，分析には商品情報マスターデータを活用する．そのマスターデータには，商品ごとに商品のカテゴリ，商品名，使用されている花材，商品の形状（“ アレンジメント ”，“ 花束 ” など）およびイメージ画像が含まれている．

3.2.1 利用する分類器

本研究では，説明変数間に交互作用が存在するため，それを交互作用ベクトルの形で表現することが可能な FM を二値分類器として用いる．ある商品 A の購買有無を推定する分類器について，目的変数 $y^A(\boldsymbol{x})$ をある商品 A を購買した場合に 1，他商品を購買した場合に -1 を取る変数，説明変数 $\boldsymbol{x}^A$ を購入用途などの購買に関する情報と顧客属性とする．なお，わかりやすさのため説明変数についても $\boldsymbol{x}^A$ と記載しているが，説明変数の値はどの商品を対象とした場合にも変化しないため，$\boldsymbol{x}$ はどの商品の購買を目的変数として扱っていても，常に一定の値を持っている．

以下の式 (3) より商品 A の特性を表現するパラメータとして，各変数の直接効果を表す係数 $\boldsymbol{w}^A = (w_0^A, w_1^A, ..., w_d^A)$ および，交互作用を表す行列 $\boldsymbol{V}^A \in \boldsymbol{\mathcal{R}}^{d\times k}$ が学習される．なお，$\boldsymbol{V}^A = (\boldsymbol{v}_1^A, \boldsymbol{v}_2^A, ..., \boldsymbol{v}_d^A)^\top$，$\boldsymbol{v}_i^A = (v_{i1}^A, v_{i2}^A, ..., v_{ik}^A)^\top$ である．

$$\hat{y}^A(\boldsymbol{x}) = w_0^A + \sum_{i=1}^{d} w_i^A x_i^A + \sum_{i=1}^{d}\sum_{j=i+1}^{d} \langle \boldsymbol{v}_i^A, \boldsymbol{v}_j^A \rangle x_i^A x_j^A \quad (3)$$

$$\langle \boldsymbol{v}_i^A, \boldsymbol{v}_j^A \rangle = \sum_{l=1}^{k} v_{il}^A v_{jl}^A \quad (4)$$

なお，$\boldsymbol{w}^A$ と $\boldsymbol{V}^A$ を学習する際に最小化すべき損失関数は，以下の式 (5) で表される．ここで，$\hat{y}^A$ は予測値を示している．

$$\ln\left(\exp(-y^A \hat{y}^A) + 1\right) + \lambda_w \|\boldsymbol{w}^A\| + \lambda_v \|\boldsymbol{V}^A\| \quad (5)$$

ただし，$\|\boldsymbol{\alpha}\|$ はベクトル $\boldsymbol{\alpha}$，もしくは行列 $\boldsymbol{\alpha}$ の 2 次ノルムを表し，λ_w と λ_v はパラメータの過学習を防ぐための正則化パラメータである．提案手法においては，交互最小二乗法を用いてパラメータの推定を行っている．

3.2.2 負例データ選択方法

対象商品 A を購買したデータを正例とし二値分類器を学習する場合，負例の数が極端に多くなり適切な係数の学習が困難になるため，負例データをサンプリングし，データ数をそろえる必要がある．ここで，二値分類器は正例と負例を分類する係数を学習するため，負例として選択されたデータにより，得られる係数も変化する．また，係数を顧客の嗜好を反映した商品特性と解釈し，その値を用いて類似性評価を行うためには，二値分類器が商品 A を好む顧客と，そうではない顧客の嗜好を反映するものでなければならない．例えば，商品 A のカテゴリが “母の日” であったとき，負例として “お盆用” や “開店祝い用” などの商品が選択された場合には，購入用途のみを用いてそれらを分類することが可能となり，そこで得られる係数には顧客の嗜好を反映することができない．すなわち，負例として選択される商品は，正例の商品 A を購買する顧客が比較対象としたものの，嗜好に合致せず好まなかった商品である必要がある．

ここで，適切な負例の選択方法を考えるにあたり，対象の生花 EC サイトのページ構成を活用する．本研究の対象生花 EC サイトでは，顧客は購入用途のカテゴリに属する商品が掲載されているページを閲覧し，複数の商品を検討したうえで，最終的に自身の嗜好に合致する商品を購入することが多い．図 3 に，対象生花 EC サイトのウェブページイメージを示す．

図 3 対象生花 EC サイト ウェブページイメージ

対象生花 EC サイトでは，顧客の購入用途となりうるイベントに合わせ，図 3 の左に示すような特集ページを作成している．そのページには，例えば “春のお誕生日” の時期であれば，“春

のお誕生日”カテゴリの商品が提示されており，それらの中で顧客は自身の嗜好に合致する商品の詳細ページを閲覧し，最終的な購入商品を決定する．つまり，同じカテゴリに属している商品であり，かつ購買されなかった商品は，顧客の嗜好に合致しなかったものと考えられる．

また，顧客は自身の興味に全く合致しない商品については，商品の詳細ページを閲覧することもなく，検討対象から外すものと考えられる．その際には，商品閲覧ページを開くことが無いため，顧客がその商品に興味を持っていなかったか否かの情報を，EC サイトの閲覧履歴データから取得することはできない．また，顧客がその商品詳細ページを閲覧しなかった理由について，嗜好に合致しないためであったのか，その他の理由があったかを知る方法は存在しない．そのため，顧客が嗜好に合致せず購入に至らなかった商品を，閲覧履歴データを用いて抽出することは困難である．従って，顧客が商品 A と比較対象とする可能性のある商品を選択する方法として，カテゴリ情報を活用することが最も適切であると考えられる．

3.2.3 係数を用いた類似度評価

二値分類器の係数を用いて，商品の類似度評価を行う手法について説明する．係数を用いた類似度評価を行う際，FM の適用のために，購買履歴データに含まれる各情報ごとに，質的変数をダミー変数を用いて変換していることに留意する必要がある．図 4 に，FM で学習される交互作用行列と，それから得られる説明変数間の交互作用を示す $d \times d$ 次元の行列について，その特徴を示す．

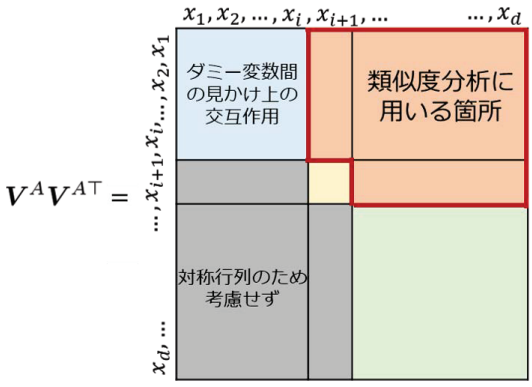

図 4　類似度評価に用いる箇所

まず，図 4 に示す $d \times d$ 次元の行列は，対角行列であるため右対角成分に着目する．ここで，FM で入力される説明変数 $\boldsymbol{x}$ はその値が 0, 1 となるよう，各質的変数ごとにダミー変数を用いて 1hot ベクトルとしている．すなわち，$d \times d$ 次元の行列中には，ダミー変数の導入により同じ質的変数間に生じた見かけ上の交互作用が存在しており，その値を類似度評価に用いることは適切ではない．従って，本研究で類似度分析に用いる箇所は，図 4 の右上の箇所のみとなる．

次に，類似度評価指標の定式化を行う．交互作用行列 $\boldsymbol{V}^A$ の積 $\boldsymbol{V}^A\boldsymbol{V}^{A\top}$ のうち，類似度分析に用いる値を並べたベクトル $\tilde{\boldsymbol{w}}^A = \tilde{\boldsymbol{w}}^A(\boldsymbol{V}^A\boldsymbol{V}^{A\top})$ と，FM で学習された直接効果 $\boldsymbol{w}^A$ を用いて，商品 A と B 間の類似度 $S(A,B)$ を式 (6) のように定義する．ただし，$\boldsymbol{W}^A = (\boldsymbol{w}, \tilde{\boldsymbol{w}})$ である．

$$S(A,B) = \frac{\langle \boldsymbol{W}^A \boldsymbol{W}^B \rangle}{\|\boldsymbol{W}^A\| \|\boldsymbol{W}^B\|} = \frac{\boldsymbol{W}^A \boldsymbol{W}^{B\top}}{\|\boldsymbol{W}^A\| \|\boldsymbol{W}^B\|} \tag{6}$$

4　実データ分析結果と考察

4.1　適切なパラメータの選択

提案手法では二値分類器として FM を用いている．FM では，交互作用ベクトルの次元数 k と，正則化パラメータ λ_w と λ_v の値を自身で設定する必要がある．そこで，年間被購買数量上位 200 位の商品を対象とし，それらが商品 A を購買したデータであったか否かを予測する問題を作成することで，適切なパラメータの選択を行った．その際の実験条件を以下に示す．

- 期間：2018 年 8 月–2019 年 7 月（注文日ベース）
- 商品：年間被購買数 上位 200 商品
- テストデータサンプリング数：各商品最大 100 件
- 学習データサンプリング数：各商品最大数
- 目的変数：各商品の購買有無
- 説明変数：購買に関する情報と商品を購買した顧客属性

なお，各商品についてその商品を購入した顧客の嗜好を学習するため，対象 200 商品それぞれについて，その商品を正例としたデータを作成している．また，その商品の購買データを正例として用いる場合，可能な限り多くのデータを学習データに用いて実験を行い，負例のデータ数はテストデータ，学習データ共に正例と同等としている．事前実験の結果より，交互作用ベクトルの次元数 $k = 10$，正則化パラメータとして $\lambda_w = 0.01$, $\lambda_v = 0.5$ を用いることとした．

4.2　某商品を対象とした場合の詳細分析事例

4.2.1 分析対象データ詳細

分析対象の購買履歴データ詳細を以下に示す．なお，期間，目的変数，説明変数については検証実験と同様である．

- 商品：年間被購買数 上位 1,000 商品
- 対象購買履歴数：各商品最大 100 件（正例時）

具体的な説明変数は購入用途，受注時間帯，購買月，購買曜日，注文日と届け日までの差（一週間ごと，13 週以上は 1 つにまとめる），性別，年代（10 歳ごと）および法人フラグであり，各属性ごとに 1hot ベクトルに変換を行っているため，最終的な説明変数の次元数 d は 102 となっている．

4.2.2 類似度分析結果

分析事例として花材に “紫リンドウ”と “トルコキキョウ”を用いた，カテゴリが “お盆”で，商品名が “お供え用のアレンジメント”の商品を対象商品とし，提案手法による類似度の分析結果を示す．なお，対象商品との類似度は式 (6) を用いて算出した．対象商品と高類似度商品の商品画像を図 5 に示す．なお，商品画像を囲む四角は，その商品に対象商品と同様の花材が用いられていることを意味している．

図 5 に示す高類似度 10 商品のうち，対象商品と同じカテゴリ “お盆”に属する商品は 5 位の商品のみであり，その他の商品は全て “お誕生日”や “開店祝い”など，異なるカテゴリの商品

図 5　分析対象商品とその高類似度商品群

であった．さらに花材に着目すると，対象商品と同様に “トルコキキョウ”を利用する商品が 5/10 含まれていた．花の種類には顧客の嗜好が現れていると考えられるため，本提案手法で算出した類似度には，顧客の嗜好が反映されているといえる．さらに，図 5 に示す実際の商品写真を見ると，全体の形状や雰囲気など，定性的な観点からも分析対象商品に似ている商品が高類似度商品として評価されていることが分かった．実際に，企業担当者の視点からも，定性的な観点について同様の結論を得ることができた．すなわち，同一カテゴリの商品を分類器の負例として学習することで，異なるカテゴリの商品を高類似性商品として評価することができており，研究目的に合致した指標を得られたといえる．

4.2.3　その他の負例選択方法との比較

本提案手法では，ある商品の購買有無を識別する二値分類器を学習するためのデータセット生成時負例データのサンプリングを行っていた．ここで，選択される負例により学習される分類器の係数が変化し，その結果類似度分析結果も変化すると考えられる．そのため，その際の負例選択方法によって類似度分析結果が異なることを示し，提案手法による負例選択方法が最も適切であることを示すために，負例選択方法のみを変化させた実験を行った．なお本研究では比較手法として，購買履歴データおよび関連情報から得られる情報を用いたサンプリング対象の負例選択を行っている．

a）　比較手法概要

具体的には，以下の商品群をそれぞれの手法における負例選択対象とする．

- すべての商品 (以下，比較手法 1)
- 商品 A を購買した顧客による購買数下位 N%の商品 (以下，比較手法 2)
- EC サイト上での商品 A との同一顧客閲覧数下位 N%の商品 (以下，比較手法 3)

以上の商品群を対象とし，負例のサンプリングを行った．

ここで，比較手法 1 は，商品 A 以外の商品を購買したすべての購買データを負例の対象とする方法である．比較手法 2 では，購買履歴データに含まれる顧客 ID を用いて，同一顧客の購買商品を抽出している．1 年間の購買商品が 2–99 個の顧客の購買履歴データを対象とし，データの傾向を考慮し，同一顧客による購買が少ない下位 $N = 60\%$ の商品を負例の対象データとして用いた．比較手法 3 は，EC サイト上の閲覧履歴に基づく負例選択を行う手法である．具体的には，同一 IP アドレスからのアクセスがあったデータを同一顧客からのアクセスであると見なし，負例抽出の基準としている．商品 A を閲覧した顧客が閲覧した商品を調べ，データの傾向を考慮し，その回数が下位 $N = 20\%$ の商品を負例の対象として抽出した．それぞれ，負例として選択される商品群のカテゴリや主な購入用途が異なっていることが確認された．

また，提案手法と同様の条件で検証実験を行い，FM のパラメータについて適切な値の推定を行った．その結果，交互作用ベクトル次元数 k について，それぞれ比較手法 1 では $k = 10$，比較手法 2 では $k = 3$，比較手法 3 では $k = 2$ とし，正則化パラメータは提案手法と同等の $\lambda_w = 0.01$ と $\lambda_v = 0.5$ を用いることとした．

b）　類似度分析結果

図 5 に示す対象商品と同じ，花材に “紫リンドウ”と “トルコキキョウ”を用いた，カテゴリが “お盆”で，商品名が “お供え用のアレンジメント”の商品を対象商品とした場合の類似度評価結果について示す．負例データの選択方法として閲覧履歴データを用い，対象商品の商品詳細ページを閲覧した顧客による，商品詳細ページの閲覧が少ない商品群を負例として選択した際の類似性分析結果について図 6 に示す．

図 6　閲覧履歴データに基づく負例選択時 類似度上位商品

図 6 に示す類似度上位 10 商品のうち，類似度 4, 5, 7, 10 位の商品のカテゴリは，対象商品と同じ “お盆”であった．この結果から，3 点の傾向を指摘することができる．まず，同じカテゴリの商品の類似度を高く評価していることがある．本研究の目的は，ある用途で購買した顧客に対し，他の用途での購買を促すための適切な商品提示を行うための，商品類似度の評価であった．カテゴリは購入用途となりうるイベントに紐づき決定されていることから，同じカテゴリに属する商品は同じ用途で購入されやすいと考えられるため，この結果は不適切であるといえる．次に花材に着目すると，異なるカテゴリに属し，かつ同じ花材を用いている商品が 1 つしか存在しないことがわかる．3 つ目に，定性的な評価からも，この負例選択方法による類似度評価結果は不適切であるといえる．

同様に，他の 2 つの比較手法を用いた場合でも，カテゴリや花材，定性的な観点から，類似度評価結果として不適切な結果が見られた．つまり，負例データによって得られる商品類似度も変化することがわかる．さらに，提案手法の負例選択方法が研究目的に対して最も適切であると考えられ，その有効性を示すことができた．

4.3　異なる商品を対象とした場合の分析事例

上で取り上げた対象商品以外についても同様の分析が行えることを示すために，いくつかの商品を抜粋し，その類似度評価結果を述べる．

4.3.1 “母の日”カテゴリ商品を対象とした分析

カテゴリが “母の日”の商品の分析事例として，花材に “カーネーション”と “バラ”を用いた，商品名 “ティーブーケ”の商品について，類似度分析結果を示す．なお，対象商品との類似度は式 (6) を用いて算出した．対象商品と高類似度商品の商品画像を図 7 に示す．なお，商品画像を囲む四角は，その商品に対象商品と同様の花材が用いられていることを意味している．

図 7　分析対象商品とその高類似度商品群 (“母の日”カテゴリ)

また，これら類似度上位 10 商品には対象商品と同じカテゴリ “母の日”に属する商品は存在しておらず，異なるカテゴリの商品を高類似度商品として抽出できていることがわかる．さらに花材に着目すると，対象商品と同様に “カーネーション”もしくは “バラ”を利用する商品が 8/10 含まれていた．また，図 7 に示す実際の商品写真を見ると，全体の形状や雰囲気など，定性的な観点からも分析対象商品に似ている商品が高類似度商品として評価されていることが分かった．従って，本提案手法で算出した類似度には，顧客の嗜好が反映されていると考えられる．

次に，異なる負例選択方法を用いて，同じ商品を対象として類似度を分析した結果と比較する．負例をランダムに選択 (比較手法 1) した際の分析結果を，図 8 に示す．

図 8　分析対象商品とその高類似度商品群 (比較手法 1)

類似度上位 10 商品にはカテゴリが “母の日”の商品は含まれておらず，異なるカテゴリの商品の類似度を高く評価することはできていた．しかし，類似度 1,6,10 位の商品は特定の花材 “バラ”のみを主とした商品であり，定性的な観点から対象商品と類似しているとは考えにくい．また，比較手法 2 と 3 の双方についても同様の分析を行った結果，類似度上位 10 商品中それぞれ 5, 6 商品が対象商品と同じカテゴリ “母の日”に属する商品であった以上より，本対象商品に対しても，提案手法による類似度分析結果が最も適切であることが示された．

4.3.2 “父の日”カテゴリ商品を対象とした分析

カテゴリが “父の日”の商品の分析事例として，花材に “オレンジバラ”を用いた，商品名 “オレンジバラのアレンジメント”の商品について，提案手法による類似度分析結果を図 9 に示す．

図 9 より，その商品の形状や色味が対象商品と似ていると考えられる．また商品情報上では対象商品の花材として “バラ”のみが記載されているが，画像を見る限り，“カーネーション”や “トルコキキョウ”も用いられている．それらの花材が用いられているかという観点からいえば，提案手法により評価された高類似度商品は同じ花材を用いており，したがって顧客の嗜好を反映した類似度評価が行えていると考えられる．

図 9　分析対象商品とその高類似度商品群 (“父の日”カテゴリ)

また，比較手法による類似度評価結果を見ると，特定の花材を用いた商品が高類似度と評価されているなど，特に定性的な観点から，適切な類似度評価が行われたとはいえない．すなわち，“父の日”カテゴリの商品についても，提案手法による類似度分析結果が適切であることが示された．

4.4 交互作用を用いることの有効性評価

本研究では説明変数間の交互作用を考慮するため，分類器に FM を用いた．交互作用の考慮による影響を明らかにするため，交互作用を考慮した場合としない場合について，各説明変数に対して商品毎に推定された回帰係数の分散を，降順で図 10 に示す．左図が交互作用を考慮しない場合，右図は提案手法の結果を表す．また，1 対 1 の比較を行うため，どちらも各説明変数の直接効果 $\boldsymbol{w}^A$ のみを示している．

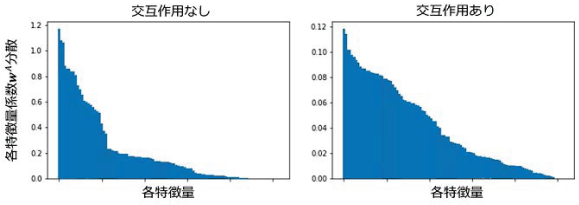

図 10　説明変数別 係数分散

ここで，分散の値が他の説明変数に比べて大きい説明変数は，商品により異なる係数の値を持つということであり，すなわち商品特性に影響を与える説明変数である．図 10 より，交互作用を考慮しないモデルで推定された係数の分散は，一部の説明変数についてのみ分散が大きくなっている．一方，交互作用を考慮することにより，比較的多くの特徴量について係数の分散が大きくなっていることがわかる．すなわち，交互作用をモデルに組み込むことにより，様々な説明変数によって商品特性をとらえることができたと考えられる．

その理由についてさらに考察を行う．特定の変数のみ係数の分散が大きいということは，それらが説明変数に与える影響が大きい，すなわち，商品購買有無との関係性が強いことを示している．具体例として，カテゴリが “お祝い”の，送別会を主な用途として購買される商品 A について考える．多くの送別会は 3 月に実施されるため，購入月に関する説明変数のうち，“3

月”の係数が大きくなることが予想される．ここで，交互作用を用いなかった場合，購入月とその他の説明変数の関係性は考慮されず，直接効果を示す係数のみが大きくなる．一方，交互作用を用いることで，購入月と性別，年代など，さまざまな説明変数との関係性をモデル化することが可能となる．すなわち，交互作用を考慮することにより，影響が大きい説明変数について，その影響を他の特徴量との交互作用の形で分割することが可能となる．その結果，多くの説明変数との関係性から商品特性を捉えることができ，分類精度の向上が見られるなど，適切な係数を学習できたと考えられる．

4.5　対象商品による類似度分析結果の違い

提案手法を用いた類似度評価と，その結果のビジネスへの応用を行う際には，本提案手法が特定の商品だけでなく，様々な商品について有効であり適用可能である必要がある．ここで本研究目的を考えると，異なるカテゴリに属する商品の類似度を高く評価することが望ましいといえる．そこで，対象商品による分析結果の違いを，高類似度商品の同じカテゴリに属する商品割合から分析を行った．その結果，多くの商品において提案手法を用いた類似度評価によって，異なるカテゴリの商品の類似度を高く評価することができていた．

一方で，提案手法による類似度分析を行った結果，高類似度と評価された商品のカテゴリが対象商品と一致しやすい商品カテゴリが存在していた．そういった特徴を持つカテゴリの例として，“開店祝い”カテゴリに属するある商品を対象に行った類似度分析結果に関して，同じ商品カテゴリに属する商品が含まれる割合を図 11 に示す．

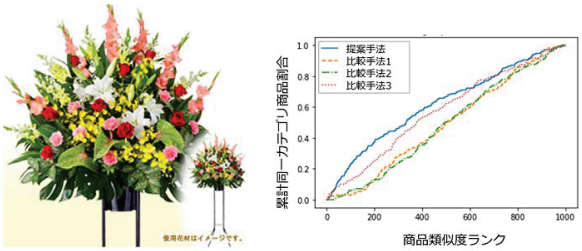

図 11　“開店祝い”カテゴリ商品を対象とした際の同一カテゴリ商品割合

左図に対象商品のイメージ画像を示し，右図に，商品類似度ランクごとに，それまでの商品に含まれる同一カテゴリ商品の割合の累計を示している．ここで，右図の縦軸が高いほど高類似度商品中に示す同一カテゴリ商品の割合が大きいことを示しており，提案手法を用いた場合に同一カテゴリ商品の類似度が高く評価されていることがわかる．この傾向は同じ “開店祝い”カテゴリに属する他の商品や，“花鉢”カテゴリに属する商品において多く見られた．

ここで，これらのカテゴリの商品は主に，店舗や企業が開業したときや開業記念日を迎えたときに祝いとして贈られるものである．その購買行動の多くは法人から法人へ送られるものであり，これらの商品を購買する法人が，その他の個人顧客が購買する商品を購入しやすいとは考えにくい．すなわち，これらの商品を購入した顧客の次の購入商品も，同じカテゴリに属する商品であると考えることが妥当である．したがって，“開店祝い”カテゴリや “花鉢”カテゴリの商品については同一カテゴリの商品の類似度を高く評価することが適切であり，その点からも提案手法による類似度分析の有効性を示すことができた．

5　まとめと今後の課題

本研究では，購入間隔が長く，購入の度に顧客の嗜好が変化するような商品を対象に，商品を購入した顧客の属性と，その商品を購入する際に顧客が検討対象とするであろう商品群の情報から，商品の類似性を評価する手法を提案した．その結果，従来手法を適用することが難しいと考えられる商品群に対しても適用することが可能であり，データを活用したマーケティングの適用可能分野を広げることができると考えられる．

今後の課題として，分析対象商品群を拡張することが挙げられる．今回は年間被購買数上位 1,000 位までの商品について分析を行っているが，実際には年間被購買数がさらに少ない商品が多数存在している．それらについても精度を高く係数を取得し適切な類似度を算出するモデルを構築することで，提案手法の適用範囲を広げることが可能となる．また，本提案手法に因られた情報を実際のビジネスに反映し，その結果を検証することにより，本手法の有効性についてさらなる検討を行うことが可能になると言える．

謝　　辞

本研究を行うにあたり用いた貴重なデータは花キューピット株式会社様よりご提供いただきました．深く感謝致します．

文　　献

[1] 経済産業省，“令和元年度内外一体の経済成長戦略構築にかかる国際経済調査事業 (電子証取式に関する市場調査),” 2020.
[2] J.K. Gerrikagoitia, I. Castander, F. Rebón, and A. Alzua-Sorzabal, “New trends of intelligent e-marketing based on web mining for e-shops,” *Procedia-Social and Behavioral Sciences*, vol.175, no.1, pp.75–83, 2015.
[3] 鷲尾和紀他，“One to one マーケティングとリレーションシップ・マーケティング: 顧客との関係性構築と one to one マーケティングの視点を中心として,” 高千穂論叢，vol.49，no.1，pp.273–311，2014.
[4] L. Ma and B. Sun, “Machine learning and ai in marketing–connecting computing power to human insights,” *International Journal of Research in Marketing*, vol.37, no.3, pp.481–504, 2020.
[5] O. Barkan and N. Koenigstein, “Item2vec: neural item embedding for collaborative filtering,” *IEEE 26th International Workshop on Machine Learning for Signal Processing*, IEEE, pp.1–6 2016.
[6] O. Barkan, A. Caciularu, O. Katz, and N. Koenigstein, “Attentive item2vec: Neural attentive user representations,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pp.3377–3381 2020.
[7] S. Rendle, “Factorization machines,” *2010 IEEE International Conference on Data Mining*, IEEE, pp.995–1000 2010.
[8] 佐和隆光，回帰分析，朝倉出版，1979.
[9] S. Rendle, Z. Gantner, C. Freudenthaler, and L. Schmidt-Thieme, “Fast context-aware recommendations with factorization machines,” *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp.635–644, 2011.