論文・事例研究

商品の潜在的類似性に基づくクラスタリング手法の提案

白井　康之，森田　裕之，後藤　裕介

1. はじめに

近年，ポイントカードの普及などにより，消費行動における ID-POS データの普及が顕著である．同一顧客が時系列でどのような商品を購入しているのかを分析することにより，将来的な顧客の嗜好を明らかにし，適切なマーケティング戦略を適用できる可能性がある．

スーパーマーケットでは，食料品や日用品といった極めて日常生活に関連の深い商品を扱っており，また，一般的に購買頻度も高いことから，特にこのような ID-POS データをもとにした分析への期待が大きい．スーパーマーケットにおける商品データは，一般に JAN コードなどの商品を特定するためのコードをもとに蓄積されている．JAN コードはバーコードとして商品に表示され，ID-POS のみならず，受発注や在庫管理などにも広く利用されている．また，会員カードなどのユーザを特定する手段と併せて，いつだれがどのような商品を購入したかがデータとして利用可能である．

一方，消費者の嗜好を解析するには，商品単位の JAN コードは分類として細かすぎることが多い．たとえば，一般的なスーパーマーケットで販売されている商品の JAN コードは数万から数十万にも及ぶ．JAN コードは容量や色の違いなどにより異なるコードが振られることが多いが，消費者の嗜好を分析するには，必ずしもこれらを区別する必要はない．すなわち，消費者の嗜好を的確に表現できるレベルの商品分類が必要である．また，JAN コードのないインストア商品のコード化方法やその分類の方法もまた重要である．

以上のように，多様な商品を取り扱う小売業の ID-POS データ分析においては，商品情報をどのようにまとめて取り扱うかが大きな課題となっている．しかしながら，商品数が膨大に及ぶうえ，類似しているという概念自体が明確に定義できるものはなく，また分析目的にも強く依存するため，こうした分析用の分類コード自体を新たに定義することは極めて困難であった．

筆者らは以前，こうした問題意識から，商品名の文字列類似性に基づく商品グルーピングの提案を行っている [1]．商品名（文字列）の類似性を編集距離により定義し，一定の閾値を超えた類似性を互いにもつ集合を商品グループとして定義した．しかし，この方法論は以下の二つの問題点をもつ．第一に，商品名がどの程度似ていれば同一商品とみなしてよいのか，一般には判定が困難である．ほぼ同じ商品でありながら，編集距離が大きく異なる商品も存在する．逆に，商品名としては類似しているものの，分析目的によっては別商品とみなしたほうがよい商品も存在する．たとえば，カレー商品の分析においては，単に「甘」と「辛」といった一文字だけの相違であっても，顧客の嗜好を判断するうえでは重要な相違である．第二に商品名の記述が必ずしも商品特性を表したものとはなっていないことである．一般に商品の特性は，商品説明文などからも抽出は可能であるが，商品説明文はメーカにより記述レベルがまちまちであり，統一的な基準を前提とした処理を行うのは困難である．

以上のような問題点を踏まえ，本研究では，商品名や商品説明文からの関連性ではなく，その商品がどのようなほかの商品と併買されているかという情報から，商品の潜在的類似性を定義しようと考えた．すなわち，他商品との共起購買パタンが類似している場合には，それらは潜在的に同じカテゴリの商品群であるとみなす．

しらい　やすゆき
大東文化大学経営学部
〒 175–8571　東京都板橋区高島平 1–9–1
yasuyuki.shirai@gmail.com
もりた　ひろゆき
大阪府立大学現代システム科学域知識情報システム学類
ごとう　ゆうすけ
岩手県立大学ソフトウェア情報学部
受付 16.7.25　採択 16.11.9

本研究では，実際のID-POSデータ[1]を用いて検証を行っているが，このデータには，複数の異なるスーパーマーケットチェーンからのデータ[2]が含まれている．一般にスーパーマーケットの購買データでは，チェーン独自の，もしくはチェーンによってかなり偏りのある商品が存在している．したがって，JANコードは付与されてはいるものの，販売店舗の偏りが大きいものや，インストア商品については，もともとJANコードが付与されていないものも存在する．

本稿では，以上のようなデータをもとに他商品の併買状況の類似性をもとにした商品分類を行う．この方法では，ある商品群の分類は，ほかにどのような商品と併買されているかによって特徴づけられるが，その意味でこの分類手法は相互依存的である．したがって，本稿では，相互に依存する商品群のクラスタリングを再帰的に繰り返すことで，最終的な分類結果を得る手法を提案する．本稿では，この手法を「ローテーションクラスタリング」と呼ぶ．本手法で得られるクラスタリング情報を用いることにより，たとえば，メーカ別ではなく，高級な肉類とよく一緒に購入されている牛乳は何か，健康志向のヨーグルトとよく一緒に購入されている牛乳は何か，といった分析が可能となる．

以下，本稿の構成は以下のとおりである．2節では，ローテーションクラスタリングの方法を示す．3節では，実データを用いた解析結果として，前述のID-POSデータを用いたクラスタリング結果を示し，クラスタリング結果の評価を行う．また，4節では，本クラスタリングの応用可能性について言及する．5節では，関連する既存研究との差異を示し，6節で，本研究のまとめならびに今後の課題を整理する．

2. ローテーションクラスタリング

今，N 個の商品群 $M_1, M_2, \ldots, M_N$ について，それぞれクラスタリングを行うことを考える．ここで M_I は，たとえば，「牛乳」などの商品グループを表す．M_I の各要素は，牛乳の各個別商品（JANコードで特定される商品単位，またはJANコードが付与されていない商品については，i-codeと呼ばれる商品コードで特定される商品単位）に対応する．商品群 M_I のクラスタリングにおいては，商品群 $M_J(I \neq J)$ との共起購買行列をもとにしてクラスタリングを行い，ここで得られたクラスタ情報をもとに，商品群 M_I と商品群 $M_J(I \neq J)$ の各商品間の共起購買行列を更新する．Step 1として，各 $M_I(I = 1, 2, \ldots, N)$ のそれぞれについてクラスタリングを行ったのち，共起購買行列を更新したうえで，Step 2として各商品群のクラスタリングを繰り返す．以上がローテーションクラスタリングの基本的な考え方である．

最大繰り返し回数を R とした際のローテーションクラスタリングのアルゴリズムを以下に示す．

準備. 異なる商品群間の共起購買行列を $P^0_{1,2}$, $P^0_{1,3}$, $\ldots$ $P^0_{N-1,N}$ とする．ここで，$P^0_{I,J}$ は，商品群 M_I,M_J 間の共起購買行列である（0は初期値であることを示す）．共起購買行列 $P^0_{I,J}$ の各要素 $m^0(i,j)$ は，商品 i $(i \in M_I)$ と商品 j $(j \in M_J)$ の併買に関するJaccard係数とする．すなわち，商品 i と j をともに購入した顧客数を商品 i または j を購入した顧客数で除したものである．

1. 対象商品群 M_I について，M_I 以外の商品群 $M_J(I \neq J)$ を説明商品群とし，商品群 M_I を $P^r_{I,J}$ の情報を用いてクラスタリングを行う（$r = 0, 1, 2, \ldots, R-1$ は繰り返し回数を表す）．ここで，クラスタリングアルゴリズムはEMアルゴリズムとし，クラスタ数は各商品群についてそれぞれ指定するものとする．本ステップでは，対象商品群を入れ替えてクラスタリングをそれぞれ行うため，合計で N 回のクラスタリングを行うことになる．

2. 各対象商品群について，クラスタリングの結果から各クラスタの重心を求め，以下の補正を行う．対象商品群を M_I としたときの共起購買行列 $P^r_{I,J}$ の各要素 $m^r(i,j)$ について，前ステップで対象商品群を M_J としてクラスタリングした際の結果に基づき，温度パラメータ ΔT_r（比率）を使って以下のように補正し，$P^{r+1}_{I,J}$ を構成する．

$$m^{r+1}(i,j) = m^r(i,j) + \Delta T_r(C^r_j(i) - m^r(i,j))$$

ここで，$C^r_j(i)$ は r 回の繰り返し後に，商品 j の属するクラスタの重心の i 番目の要素を表す．

3. 1, 2の手順を R 回繰り返す．

[1] 経営科学系研究部会連合協議会による平成27年度データ解析コンペティションにおいて，（株）アイディーズより提供いただいたデータ．

[2] チェーンについては，複数のチェーンストアがデータ中に存在することは情報として提供されていたが，チェーンの識別子は提供されていない．しかし全データから，顧客の店舗併用パタンをグラフで表現したところ，複数のサブグラフに完全に分離していることがわかった．これは，そのサブグラフ内でのみ顧客IDが共通になっていることの証左であると考え，以下では，このサブグラフを一つのチェーンと呼ぶことにする．

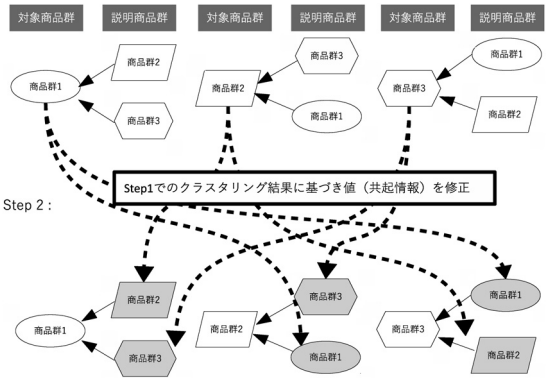

図 1　ローテーションクラスタリングの概念図

ここで，温度パラメータ ΔT_r は，繰り返し回数 r を用いて，以下のように定義される．

$$\Delta T_r = v\Delta T_{r-1} \quad (r \geq 1)$$

v は繰り返し回数による温度の減少率を表すもので，以下，減衰係数と呼ぶ．たとえば，ΔT_0 を 0.2 に，また，v を 0.8 とすれば，ΔT は 0.2, 0.16, 0.128, 0.102, 0.082 のように変化する．

温度変化の大きいときは，クラスタリング結果にも影響を与えうるが，温度変化が 0 に近くなると，クラスタリング結果は影響を受けにくくなる．特に，温度変化が 0 に限りなく近づいた場合には，説明商品群との共起購買行列が全く変化しなくなるので，クラスタリング結果は変化しない（すなわち収束する）．

ローテーションクラスタリングの概念図を図 1 に示す．図のように対象商品群をローテーションしながらクラスタリングを繰り返す．Step 1 で得られたクラスタ情報をもとに，共起購買行列を補正し，Step 2 のクラスタリングを繰り返す．

ローテーションクラスタリングの簡単な例を図 2 に示す．たとえば，牛乳 2 とトマト 1 の共起は当初は 0 であったが，他商品でのクラスタリング結果を反映することにより，Step 2 では 0.038 という値になっている．これは，牛乳 2 を販売する店舗では，トマト 1 が売られていなかったために，初期段階では共起が 0 になっているが，仮にこの店舗でもトマト 1 を扱うとしたならば，他商品の購買状況から見て，牛乳 2 とトマト 1 の共起が発生するであろうという直観とも一致している．

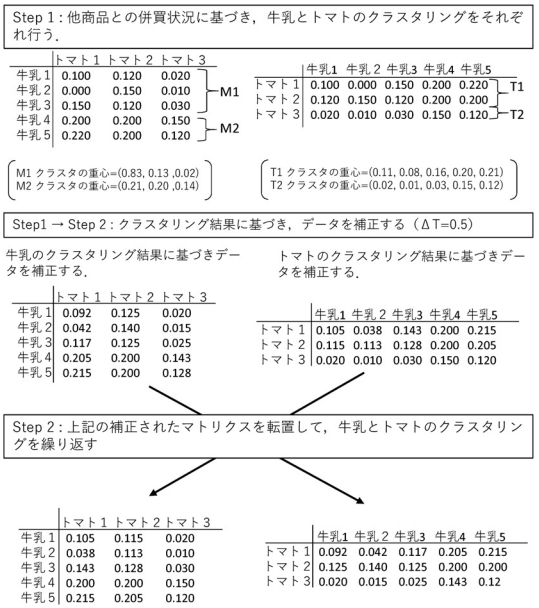

図 2　ローテーションクラスタリングの例

3.　実データを用いた実験と評価

3.1　実験データ

本研究で対象とした実験データは，複数のチェーンにおけるスーパーマーケットの購買履歴（2013 年 7 月～2015 年 6 月まで）からなる．使用したデータ項目は以下のとおりである．

・ID-POS データ
　購入者，購入日，購入店舗，購入金額，商品を特定する情報
・商品マスタ
　商品情報，商品分類情報，商品コード（JAN コード），商品名

本分析では，上記データのうち，二つのチェーンに属する合計 90 店舗のデータを抽出して実験を行っている．チェーンはそれぞれ便宜的にチェーン 1，チェーン 2 とする（チェーン 1 には 24 店舗が含まれ，チェーン 2 には 66 店舗が含まれる）．チェーン 1 は 246,371 ユーザが含まれ，一人当たりの購買額の平均は 277,666 円であった．また，チェーン 2 は 122,575 ユーザが含まれ，一人当たりの購買額の平均は 468,042 円であった．

汎用商品については，チェーン 1, 2 ともに同様の購買傾向があるものの，一部の商品については，チェーン 1 もしくはチェーン 2 でのみ販売されているケース，また，2 チェーンでの取り扱いはあるものの購買実績は顕著に偏っているケースなどが見られた．また，

チェーン間の相違だけではなく，店舗によって恒常的に取り扱っている商品もあれば，売上実績から見て明らかに取り扱いに偏りのあるケースも見られる．本分析の目的は，こうしたチェーンもしくは店舗による購買傾向の偏りを相殺し，購買パタンの類似性に基づき，商品の分類を行うことである．

以下の分析では，商品群として，牛乳，牛肉，ヨーグルト，トマトの四つを取り上げ，これらの併買関係から各商品群をクラスタリングすることを試みた．このうち，牛肉，トマトは，その商品の特質上，特に人気商品においてはインストア商品が多く，もともと JAN コードが付与されていない．一方，牛乳とヨーグルトは，通常は，JAN コードが付与されている汎用の商品であるが，一部，特定の店舗でのみ，もしくは特定の店舗での扱いが極端に多い商品が存在している．

ここで，共起購買行列の各要素 $m^0(i,j)$ は 2 節で示したように初期値としての商品 i と商品 j の Jaccard 係数である．商品購入の共起は厳密に考えると明確な定義を与えることは困難であり，同一レシートでの同時購入では少なくとも本分析で対象とする併買という概念から見れば狭すぎると思われたため，本分析では，分析期間中を通じて商品 i と商品 j の購入があるか否かに着目した．すなわち，ここで用いた Jaccard 係数は，分析期間中に商品 i と商品 j を（同一レシートとは限らず）購入した顧客数を，商品 i または商品 j を購入した顧客数で除したものとした．

なお，本データの分析においては，日常的に当該スーパーチェーンを利用しているユーザに限定するため，12 回以上取引が存在するものに限定した．また，本データには，マイナスの売上がたっているキャンセルデータが存在していることから，マイナスの売上データについては，顧客番号，商品情報が一致し，かつ定価ならびに日時が一定範囲内にある直近の取引数量を相殺した．

3.2 実験環境

本実験は，インテル Core i5-3320M プロセッサ (2.60 GHz)，主記憶 (RAM) 16 GB の Windows 7 PC を使用して行ったものである．クラスタリングは weka 3.6.9 [2] に含まれる EM アルゴリズムを使用した．

使用したデータは，牛乳が 105 商品，牛肉が 194 商品，ヨーグルトが 505 商品，トマトが 311 商品である．このため，共起購買行列のサイズは，たとえば，牛肉に関しては 194 行× (105+505+311) 列= 194 行× 921 列となる．

実行時間は，1 ステップ当たり平均 73 秒であり，ここには，四つの商品群のクラスタリング，行列の値補正などの処理が含まれる．

後述するように，実質的には 4 回の繰り返しでクラスタリングは収束しているが，収束状況を確認するため 30 回の繰り返しを実行している．このため，全体の実行時間は，およそ $73 \times 30 = 2,190$ 秒であった．実際には，次節で述べるように，クラスタリングの収束状況に基づき，繰り返し回数を制御するのが適当であると考えられる．

3.3 実験パラメータ

上述のデータを利用して，共起購買行列を作成し，ローテーションクラスタリングを実行した．ここで，最初の温度 ΔT_0 を 0.3 とし，減衰係数 v を 0.9，求めるクラスタ数は各商品とも 4 とした．ΔT_0 を大きくすると，初期のクラスタリングにおいて大規模なデータの書き換えが発生しやすい．多くの商品がインストア商品であるようなケースでは，ΔT_0 を大きくするほうが適切なケースもありうるが，本研究で対象としたデータでは，牛乳，ヨーグルトはほとんどが JAN コードが付与されている汎用商品であることから，比較的小さな初期値を利用することとした．なお，0.3 以外に，0.2 や 0.4 で実施した場合でも最終的なクラスタリングは大きくは変化しない．

また，減衰係数 v が 0 に近くなると，実質データの書き換えが発生しないため，クラスタリング結果は変化しない．今回の実験においては，4 回の繰り返し計算後，すべての商品群におけるクラスタリング結果は安定した．図 3 は，各商品群におけるクラスタリングの収束状況を示したものである．ここで，横軸は繰り返し回数を，また縦軸は，クラスタに属する各データとクラスタ中心との距離の総和を示したものである．距離の総和が小さくなるほど，クラスタリングは安定し，変化しにくくなっていることを表している．本図から明らかなように，各データとクラスタ中心との距離の総和は繰り返し回数を経るにしたがって小さくなり，したがってクラスタリング結果が変化しなくなっていることが想像できる．

3.4 クラスタリング結果

表 1 は，牛乳，牛肉，ヨーグルト，トマトの各商品に関するクラスタリングの結果を示したものである．

ここで得られたクラスタは，単に製造元やブランド，金額といった外形的な特徴で分類されたものではなく，併買状況からクラスタリングされたものであり，各クラスタにはそれぞれの特徴を見ることができる．たとえば，ヨーグルトの各クラスタはそれぞれ表 2 に示し

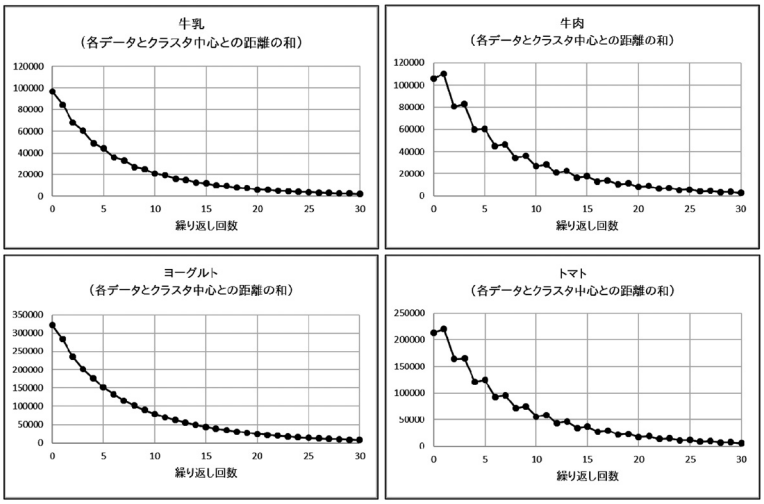

図 3　各商品群のクラスタリングの収束状況

表 1　生成されたクラスタ

商品種別	クラスタ	商品数
牛乳	G1	22
牛乳	G2	23
牛乳	G3	23
牛乳	G4	36
牛肉	B1	37
牛肉	B2	34
牛肉	B3	65
牛肉	B4	57
ヨーグルト	Y1	97
ヨーグルト	Y2	129
ヨーグルト	Y3	161
ヨーグルト	Y4	117
トマト	T1	123
トマト	T2	74
トマト	T3	34
トマト	T4	79

たような特徴をもつと解釈できた．

4.　応用可能性

一般に，購買分析においては，どのような顧客がどのような商品を購入し，また，将来的にどのような商品を購入する可能性があるかを分析することが重要であるが，特に，スーパーマーケットの購買分析においては，その膨大な商品数や取扱経路の煩雑さから，分析目的に合致した適切な粒度の商品分類が存在せず，多くの分析においてボトルネックとなっていた．

本稿で示したようなクラスタリング手法は，他商品の購買状況のみから類似した商品の分類を行うので，

表 2　ヨーグルトのクラスタ

商品種別	クラスタ	主な特徴
ヨーグルト	Y1	朝に食べる小容量のヨーグルト
ヨーグルト	Y2	大人数家族向け，大量消費のヨーグルト
ヨーグルト	Y3	付加的な要素の入ったヨーグルト
ヨーグルト	Y4	健康マニア向けのヨーグルト

汎用性が高く，また，消費者の嗜好を解析する点においても応用可能性が高い．

以下では，本手法の将来的な応用可能性について論じる．

4.1　併買可能性の評価

ローテーションクラスタリングにおいては，最終的に得られるクラスタだけでなく，書き換えられた行列の値自体もまた将来的な併買可能性を示す重要なものである．ここでは商品 i と j の将来的な併買可能性を以下の指標で評価する．

$$E_{i,j} = log(S_i) \times (m^r(i,j) - m^0(i,j))$$

ここで，r は繰り返し回数，S_i は商品 i の購入者数を表す．一般に，購入者数は商品により著しく異なっているため，実数としての差をそのまま適用すると，過度に現状売上数の多い商品のみが高い値となる可能性があるため，対数をとることとした．$E_{i,j}$ は，商品 i を購入する人が商品 j を購入する将来的な可能性を定量的に示したものである．事業者であるスーパーマーケットとしては，この値が大きいほど，売上増につながるポテンシャルが高い商品の組み合わせであるとい

表 3　併売での売上を見込むことができる組み合わせ

No.	商品1	商品1のクラスタ	商品1の販売店舗数	商品2	商品2のクラスタ	商品2の販売店舗数	評価値(E)	評価値ランク	現状の共起数	m^{10}	m^{0}
1	牛乳2(酪農)	G4	66	ヨーグルト3(果肉)	Y2	89	1245	38	1781	95.0	8.9
2	牛乳2(酪農)	G4	66	ヨーグルト1(果肉)	Y2	89	1216	39	1382	90.9	6.9
3	牛乳2(酪農)	G4	66	ヨーグルト2(果肉)	Y2	89	1188	48	1701	90.7	8.6
4	ヨーグルト5(汎用)	Y2	90	牛乳1(汎用)	G4	70	1143	57	2439	126.1	40.2
5	ヨーグルト5(汎用)	Y2	90	牛乳3(汎用)	G4	61	1140	58	2506	126.2	40.5
6	ヨーグルト5(汎用)	Y2	90	牛乳5(高級)	G4	63	1093	67	3227	133.2	51.0
7	ヨーグルト10(汎用)	Y2	90	牛乳1(汎用)	G4	70	1070	74	1844	116.1	34.8
8	ヨーグルト8(汎用)	Y2	90	牛乳1(汎用)	G4	70	1070	75	1418	135.2	49.8
9	ヨーグルト7(健康)	Y2	90	牛乳1(汎用)	G4	70	1048	83	1122	121.7	38.2
10	ヨーグルト8(汎用)	Y2	90	牛乳4(高級)	G4	80	1044	84	1729	139.1	55.7
11	ヨーグルト6(健康)	Y2	89	牛乳1(汎用)	G4	70	1043	85	1208	130.4	46.5
12	ヨーグルト5(汎用)	Y2	90	牛乳4(高級)	G4	80	1000	102	3753	134.6	59.4
13	ヨーグルト4(健康)	Y2	90	牛乳1(汎用)	G4	70	992	112	1525	125.9	47.5
14	ヨーグルト9(健康)	Y2	90	牛乳1(汎用)	G4	70	980	123	1466	122.9	45.6
15	ヨーグルト5(汎用)	Y2	90	トマト1(地場産)	T3	59	973	126	1285	94.5	21.4
16	ヨーグルト6(健康)	Y2	89	牛乳4(高級)	G4	80	973	127	1654	136.2	58.0
17	ヨーグルト5(汎用)	Y2	90	トマト2(地場産)	T3	63	970	130	1336	95.1	22.2
18	ヨーグルト10(汎用)	Y2	90	牛乳3(汎用)	G4	61	968	133	2524	120.2	46.7
19	ヨーグルト8(汎用)	Y2	90	牛乳3(汎用)	G4	61	967	134	1843	139.5	62.3
20	ヨーグルト9(健康)	Y2	90	トマト1(地場産)	T3	59	960	140	313895	109.5	33.6

表 4　他チェーンでの実測値と m^{10} との比較

	チェーン1,2 での値		他チェーン（チェーン1,2 以外）での実測値（ m^0 ）															他チェーンでの m^0 の値の分布				
No.(*)	m^{10}	m^{0}	チェーン3	チェーン4	チェーン5	チェーン6	チェーン7	チェーン8	チェーン9	チェーン10	チェーン11	チェーン12	チェーン13	チェーン14	チェーン15	チェーン16	チェーン17	他チェーンでの m^0 の平均 (A)	（チェーン1,2の）m^0 未満	（チェーン1,2の）$m^0 \sim m^{10}$	（チェーン1,2の）m^{10} 以上	$\frac{A - m^0}{m^{10} - m^0}$
1	95.0	8.9	-	-	-	-	-	-	-	-	-	-	77.6	-	-	-	-	77.6	0	1	0	**0.798**
2	90.9	6.9	-	-	-	-	-	-	-	-	-	-	56.4	-	-	-	-	56.4	0	1	0	**0.589**
3	90.7	8.6	-	-	-	-	-	-	-	-	-	-	70.1	-	-	-	-	70.1	0	1	0	**0.750**
4	126.1	40.2	-	-	-	200.6	79.1	-	-	-	-	202.4	-	130.7	118.9	134.5	73.2	134.2	0	3	4	**1.094**
5	126.2	40.5	104.4	149.7	-	-	-	90.3	-	-	-	-	-	187.3	-	-	-	132.9	0	2	2	**1.078**
6	133.2	51.0	96.6	143.3	159.5	159.6	-	125.2	78.5	97.8	-	89.2	105.9	184.4	119.9	133.6	99.7	122.5	0	8	5	**0.870**
7	116.1	34.8	-	-	-	80.8	86.3	-	-	-	-	218.6	-	72.7	102.7	88.2	83.7	104.7	0	6	1	**0.860**
8	135.2	49.8	-	-	-	77.0	91.7	-	-	-	-	225.0	-	111.2	116.5	179.8	-	133.5	0	4	2	**0.980**
9	121.7	38.2	-	-	-	-	83.7	-	-	-	-	174.1	-	98.4	113.3	-	61.7	106.2	0	4	1	**0.815**
10	139.1	55.7	118.2	-	133.8	-	-	-	-	60.9	-	-	99.9	110.6	-	106.8	-	105.0	0	6	0	**0.591**
11	130.4	46.5	-	-	-	166.9	81.8	-	-	-	-	227.5	-	106.1	123.6	154.7	77.5	134.0	0	4	3	**1.043**
12	134.6	59.4	92.0	-	130.0	104.9	-	-	-	-	-	-	113.0	137.4	-	92.0	-	111.5	0	5	1	**0.693**
13	125.9	47.5	-	-	-	119.6	-	-	-	-	-	212.1	-	117.8	140.8	114.2	137.6	140.3	0	3	3	**1.184**
14	122.9	45.6	-	-	-	93.9	104.4	-	-	-	-	219.0	-	85.9	129.9	97.7	109.8	120.1	0	5	2	**0.963**
15	94.5	21.4	-	-	118.6	-	-	-	-	-	-	-	-	-	-	-	-	118.6	0	0	1	**1.330**
16	136.2	58.0	115.2	-	114.1	103.4	-	-	-	-	-	-	121.4	132.4	-	125.7	-	118.7	0	6	0	**0.776**
17	95.1	22.2	-	-	-	-	-	-	-	-	109.1	-	-	-	-	-	-	109.1	0	0	1	**1.192**
18	120.2	46.7	126.0	124.0	-	-	-	80.4	-	-	-	-	-	154.7	-	-	-	121.3	0	1	3	**1.014**
19	139.5	62.3	136.8	118.2	-	-	-	87.4	-	-	-	-	-	165.6	-	-	-	127.0	0	3	1	**0.837**
20	109.5	33.6	-	-	51.6	-	-	-	-	-	-	-	-	-	-	-	-	51.6	0	1	0	**0.237**
行番号(No.)は，表3 の行番号と対応する.																		（合計）	0	64	30	

える．

一般に，評価値 $E_{i,j}$ が大きいものは，インストア商品など，店舗などでの独自商品が多く，このため他チェーンでの購買の実績も存在しないことが多い．しかしながら，本研究で分析対象としたチェーン 1, 2 では多くの併売はないものの，他チェーンにおいては，日常的に併売が行われているものも存在し，これらについては実績値との比較も可能である．

表 3 は，各商品の組み合わせについて，評価値 $E_{i,j}$ の大きい順に並べたものの中から，他チェーンでの購買がデータとして存在するもののみを（上位から 20 件）抽出したものである．ここで，「評価値」は上記の $E_{i,j}$ を，また，m^{10} は繰り返し回数 10 回後の 2 商品の Jaccard 係数を表す．m^0 は，2 商品の Jaccard 係数の初期値である（Jaccard 係数は見やすさを考慮し，それぞれ 10,000 倍している）．

表 4 は，表 3 の 20 の商品の組み合わせについて，本分析で対象としたチェーン 1, 2 以外の単独チェーンに

おける併買の Jaccard 係数（すなわち m^0 に対応するもの），各チェーンでの m^0 の値の分布，平均値，また，実際に他チェーンで観測された m^0 の平均が，チェーン 1, 2 における m^0 と m^{10} のどのあたりに位置するか ($\frac{A-m^0}{m^{10}-m^0}$) を示したものである．

通常，スーパーマーケットの購買分析では，他チェーンでの購入状況を参照することはできないので，結果の有用性や妥当性を検証すること自体困難であるが，本分析では他チェーンの情報も提供されていたため，実際の値に基づく比較検証が可能である．なお，表中“–”となっている箇所は，該当チェーンでの併売がなく，データとして存在しないものであることを表している．

ここで，表 3 の 1 行目 (No. 1) における“牛乳 2（酪農）”は，購買された店舗数が 66 となっており，チェーン 2 の独自商品であると思われる．また，“ヨーグルト 3（果肉）”は購買された店舗数が 89 であり，ほぼ全店舗で展開されているが，各店舗での購入状況を見る限り，極めて偏った購入状況が確認されている．この商品の組み合わせの評価値が高いということは，仮にこの 2 商品が理想的に併売された場合には，期待される Jaccard 係数が現状に比して大きくなり（すなわち同時購入の可能性が高まり），またもともとの売上規模から見ても，全体に与える影響が大きいことを示唆している．この商品の組み合わせについては，表 4 で見られるように，チェーン 13 で実際に併売が行われているが，そこでの Jaccard 係数は 77.6 となっており，m^{10} の値 (95.0) と近くなっていることがわかる．

これ以外のデータでも，評価値ランクが上位の商品の組み合わせは，チェーン 1, 2 においては比較的併売の機会が少ない商品であったと考えられるが，一方，表 4 から別のチェーンでの実際の併売状況を見ると m^{10} の値に近くなっているものが多く，本分析の結果として求められた併買パタンが実際に実現可能な値と近くなっていることが推測できる．

実際，表 4 から明らかなように，併売が行われている単独チェーンでの初期値 m^0 は，チェーン 1, 2 の m^0 を常に上回り，その値の平均は，m^{10} に非常に近いものとなっている（すなわち，他チェーンでの m^0 の平均を A としたとき，$\frac{A-m^0}{m^{10}-m^0}$ の値は 1 付近のものが多い）．これより，チェーン 1, 2 における予測値としての m^{10} は，実際に併売が多く行われている他チェーンの実測値に非常に近いものとなっており，本稿におけるアプローチの妥当性を示すものと考えている．

一方，得られた各クラスタ間の（チェーン 1, 2 における）現状での顧客数ベースの購買共起率を表 5 に示す（共起率の高い順に上位 15 件）．表 3 では，クラスタ G4 の商品とクラスタ Y2 の商品間のいくつかで，将来的な売上見込みが非常に大きいことが示されているが，実際，クラスタ G4 とクラスタ Y2 の商品は，現状でも高い共起率となっており，購買嗜好として非常に近寄ったものであることが想定できる．すなわち，表 3 で示した個別商品の組み合わせは，現状では共起が多くはないものの，同クラスタに属する他商品の併売状況を見ると，商品の仕入れや販売方法の工夫により，将来的な併買の可能性が高まることは一定の説得力をもつと思われる．

表 5　クラスタの組み合わせにおける現状の共起率

クラスタ1	クラスタ2	共起率(*1)	共起率(*2)	共起率(*3)
G4	Y2	0.789190	**0.861164**	**0.904239**
G4	T4	0.548030	0.586947	**0.892071**
T4	Y2	0.536843	**0.854932**	0.590647
T4	Y1	0.518516	0.557861	**0.880267**
T3	Y2	0.408016	**0.888302**	0.430081
B2	T3	0.407080	**0.956623**	0.414736
G4	T3	0.403775	0.420263	**0.911440**
B2	T2	0.397137	0.492203	0.672792
G4	Y1	0.369285	0.382147	**0.916472**
G3	B2	0.368894	0.587069	0.498149
B1	T4	0.332361	**0.908138**	0.343924
T3	Y3	0.317551	0.367610	0.699876
G3	T3	0.312717	**0.885781**	0.325856
G3	T2	0.311378	0.442160	0.512845
B1	Y1	0.301972	0.620054	0.370534

(*1) 共起数/クラスタ1 または2 の購買者
(*2) 共起数/クラスタ1の購買者
(*3) 共起数/クラスタ2の購買者

4.2　予測シミュレーションへの応用

また，本稿で示したクラスタリング手法は，たとえば顧客行動の全体を捉えたうえでのスーパーマーケット店舗における商品導入効果の予測シミュレーションにおいても活用することが期待できる．先行研究では，店舗内での購買履歴データから顧客の回遊行動 [3] や欠品時行動 [4] が推定されてきたが，いずれも単一店舗内での顧客行動を捉えたもので，地域内の複数店舗にまたがる顧客行動の全体を捉えたものではない．

このクラスタリング手法を活用することにより，地域内の特定店舗で未販売商品を販売したときの顧客の購買行動への影響が推定可能になり，ID-POS データの情報を使って一定期間内の顧客による各商品の購買数量を計算したものを合わせて分析することで，訪問店舗や来店回数の変化や各店舗での商品購入数量への影響が予想可能になると考えられる．

5. 関連研究

本稿の手法を，他商品との共起状況に基づき共起データの欠損を補間する方法と見れば，潜在的意味解析（以下，LSI）は，潜在的に類似しているものを結果的に同じデータに射影することから，本稿の提案と関連する技術であると考えられる．しかしながら，本稿で示した手法は，他商品との関連性において嗜好が似ているといった観点からの補間，より具体的には，他商品の暫定的なクラスタリング結果に依存してデータ補間を逐次的に行うのに対し，LSI では主成分分析に基づいた類似性に基づいてデータを近似するものであり，データ補間の目的ならびに方法が異なっている．

その他，単に商品のクラスタリング自体が目的であれば，単純に商品の購入者を説明変数としてクラスタリングする方法もありうる．ただし，この方法では，そもそも販売店舗が異なるインストア商品は完全に購入者が異なるため，仮に同じ嗜好をもった商品であっても同じクラスタにはなりにくい．また，ユーザベースのソーシャルフィルタリングによる情報推薦では，類似したユーザの購入実績に基づき商品の推薦を行うことができるが，商品の分類そのものを目的としたものではない．ソーシャルフィルタリングは，一般に本稿で示したような「商品の潜在的な類似性」に着目したものではないため，ユーザの購入に至る嗜好性を考慮した情報推薦を行うことはできない．本稿で示したクラスタリング手法の情報推薦への応用については今後の検討課題としたい．

また，温度パラメータを用いたアニーリングによるクラスタリング手法としては，たとえば，量子アニーリング法を用いたクラスタ分析手法 [5] が提案されている．ただし，この方法は，クラスタリングが局所解に陥るのを避けるための技法として提案されているものであり，本稿で示したような未知の購買可能性を補間するための方法とはその目的が異なっている．

上記以外に，関連する手法としては，複数の関連性に基づくクラスタリング手法 (multi-relational clustering) [6, 7] がある．これらの手法は，複数の多様な説明変数が存在する際に，適切に属性を選択することによって，目的に沿ったクラスタリングを行う方法論を提供するものである．スーパーマーケットの販売履歴のようなデータでは実際，上記手法で前提とするような多様な関連性を確認することもできるが，本稿で扱うクラスタリング手法は，それらのうち，併買行動に焦点を当てたものであり，前提自体が異なっている．

以上のように，本稿で示したような複数の商品群の併買状況に基づき，相互依存的な関係に基づき商品を段階的にクラスタリングする方法（あるいは同じ目的をもった類似の手法）は，筆者らの知る限り存在しない．一般的には，アイテムの特徴を示す適切なデータが存在する限り，通常のクラスタリング手法が十分に有用であるが，スーパーマーケットで取り扱う商品では，商品の特徴を示す十分な情報が得られないため（もしくはデータ化するためには非常に手間がかかるため），本手法のように併買状況のみから分類を行う方法は極めて有益であると考える．

6. おわりに

スーパーマーケットなどの小売業における販売データを解析する際，消費者の嗜好やその関連性を解析するためには，個別商品単位では明らかに細かすぎ，適切ではないことは明らかである．逆に，JAN コードより粗い分類では，商品特性までを考慮することができず，これもまた，消費嗜好を解析するうえでは十分ではない．

本稿では，スーパーマーケットで販売される商品のように，商品特性が必ずしも明確に定義されていない商品について，商品特性ではなく，ほかの商品群の併買状況から繰り返し計算によってクラスタリングを行う方法論を示し，実際のデータを用いた実験を行った．クラスタリング結果の評価は一般的に困難ではあるが，最終的なクラスタを生成する共起購買行列と初期の共起購買行列との差分により，将来的に併売可能性の高い商品の組み合わせを抽出することが可能であり，実際，実験した例においては，他チェーンにおいて収束値に近い併売状況となっていることが確認できた．

スーパーマーケットにおける商品の販売は，実際には販売チャネルの確保などのさまざまな制約があるため，一般には本分析結果をそのまま適用できるとは限らない．特にインストア商品は，取引上の制約や地域的な制約があり，簡単に併売できるわけではないことは明らかであるが，本分析結果に基づき類似の商品を検討するなど，各スーパーマーケットの商品販売戦略において有益な指標となるものと考えている．

本研究に関する今後の課題は以下のとおりである．

- クラスタリングにおけるパラメータ（クラスタ数，初期の温度パラメータ ΔT_0 と減衰係数 v）と生成されるクラスタの関係性を明らかにし，目的に応じたパラメータ設定の指針を示す必要がある．
- 本稿では，クラスタ数は各商品の特性を考慮して

4としていたが，これらは分析の目的や商品群の個別の状況によっても異なる．一般にクラスタ数を決める基準を事前に明確に決定することは困難であり，また，クラスタリングの過程でクラスタ数を最適化することも必ずしも本分析の目的からすると適切ともいえない．このため，たとえば，分析者があらかじめ商品分類のキーとなる特徴情報を指定することで，より分析目的に沿ったクラスタ数を動的に決定できるような仕組みも検討の必要がある．

・本稿ではクラスタリング手法としてEMアルゴリズムを使用しているが，どのようなクラスタリング手法がよいのかは分析目的にも依存する．また，クラスタリング手法に応じたアルゴリズムの効率化も可能であろう．たとえば，本稿のように，修正されたデータでクラスタリングを繰り返す場合には，前ステップでの結果をうまく活用することで，計算時間そのものを短縮化することも可能であると思われる．

謝辞　本分析は，経営科学系研究部会連合協議会主催「平成27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て行ったものである．関係各位に謝意を表する．

参考文献

[1] 中元政一，高嶋宏之，S. Cheung，白井康之，森田裕之，“商品分類と定価推定に関する商品特性の分析方法，”経営システム，**25**(3), pp. 164–170, 2015.

[2] Weka 3, Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/

[3] 藤野俊樹，北澤正樹，山田隆志，高橋雅和，山本学，吉川厚，寺野隆雄，“スーパーマーケットで客はどう動く？―顧客動線分析とエージェントベースシミュレーションからわかること―，”計測自動制御学会第5回社会システム部会研究会資料，pp. 57–68, 2014.

[4] 松村直樹，和泉潔，山田健太，“POSデータに基づく欠品時の顧客行動を考慮した小売店舗の購買シミュレーション，”人工知能学会論文誌，**31**(2), F-F13_1-8, 2016.

[5] 田中宗，栗原賢一，宮下精二，“量子アニーリング法を用いたクラスタ分析，”科研費特定領域研究「情報統計力学の深化と展開」研究会「情報統計力学の広がり：量子・画像・そして展開」，2009.

[6] X. Yin, J. Han and P. Yu, “Crossclus: User-guided Multi-relational Clustering,” *Data Mining and Knowledge Discovery*, **15**(3), pp. 321–348, 2007.

[7] X. Cai, F. Nie and H. Huang, “Multi-view k-means clustering on big data,” In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pp. 2598–2604, 2013.