4C1-5

グラフ特徴量を用いた識別モデルによる内在的購買行動の抽出

Logistic model for loyal customers using graph features

中原 孝信 *1　　羽室 行信 *2
Takanobu NAKAHARA　Yukinobu HAMURO

*1 専修大学 商学部
School of Commerce, Senshu University

*2 関西学院大学 経営戦略研究科
Institute of Business and Accounting, Kwansei Gakuin University

In retail stores, the loyal customers, who have a store loyalty, are widely recognized as key influences. Therefore, finding the consumers' purchase intentions for loyal customers are important for working out a strategy of the store. In conceptualising store loyalty, we have defined a repeat visiting behavior in a supermarket. The use of such behavioral measures in loyalty research is still popular. In this paper, we have made a regression model for loyal customers in the supermarket using not only association rules but also graph features which are calculated by using a NetSimile method for each purchase item. The graph features show a structural relationship between purchase items for each customer.

1. はじめに

近年の小売業界では，同じ商品や同じカテゴリを扱う店舗が多く存在し，顧客にとって店舗をスイッチするコストは低くなっている．したがって， 顧客を維持するための活動はより一層重要になり，継続的に購入する顧客を確保したり，ストアイメージを高めたりすることが店舗戦略を考える上で必要不可欠である．そして，特定の店舗に対するストア・ロイヤルティを向上させ，その度合いを高めることが重要な課題になる．これまでのストア・ロイヤルティに関する研究では，価格，立地，品揃え，店員の知識などがストア・ロイヤルティを向上するための重要な要因であるとされている [清水 1996]．また，ストア・ロイヤルティを測定するための尺度は，来店回数，競合店舗の中で最も購買金額の高い店舗，直近に購入した店舗，継続的な購買を行っている店舗などであり，これらの尺度の中では来店回数が最も多く用いられている [清水 2004]．

本研究でもストア・ロイヤルティの尺度として来店回数を利用し，ある店舗の来店回数から定義した優良顧客の判別を目的としたロジスティクモデルを構築する．また，説明変数としては，データの表層的な関係性に着目した購買商品数や相関ルールなどの特徴量だけではなく，商品間のネットワーク構造を対象にした特徴量を捉えることで，商品の購買に関する内在的な購買行動と優良顧客の関係をモデル化する．本研究で利用するデータは，経営科学系研究部会連合協議会が主催する平成 26 年度データ解析コンペティションで提供された．

2. 手法

顧客の内在的な購買行動をモデル化するにあたって，まず，顧客の購買行動を商品類似度グラフによって表現する．商品類似度グラフとは，商品を節点とし，同時購買する傾向の強い商品間に枝を張った一般無向グラフのことである．本研究では，顧客の内在的な購買行動が，商品類似度グラフの構造に現れていると仮定し，商品類似度グラフから様々な特徴量を抽出することを試みる．グラフ特徴量としては，NetSimile[Berlingerio 2012, Berlingerio 2013] と構造同値 [Sailer 1978] を利用することにした．そして，得られた特徴量を説明変数として，優良顧客を目的にロジスティク回帰モデルを構築する．

連絡先: 中原孝信，専修大学，川崎市多摩区東三田 2-1-1，
TEL:044-911-0559，nakapara@isc.senshu-u.ac.jp

2.1 商品類似度グラフ

顧客の購買行動を商品類似度グラフによって表現する．商品類似度グラフとは，ある一人の顧客についての購買履歴から，互いに類似した商品に枝を張った一般無向グラフのことである．商品間の類似度は同時購買頻度に基づいて定義する．ある顧客の来店日 $t = 1, 2, \ldots, N$ に購入した商品集合を I_t とすると，その顧客の商品購入データベースは $D = \{I_1, I_2, \ldots, I_N\}$ で表される．ここで， 2 つの商品 u, v の類似度 $sim(u, v)$ は，式 (1) で定義される．

$$sim(u, v) = \frac{|\{I_t | u, v \in I_t\}_{t=1}^{N}|}{|D|} \tag{1}$$

そしてある顧客が購入した全商品を節点集合 V とすると，任意の 2 つの商品 $u, v \in V$ について，$sim(u, v) \geq \sigma$ を満たすような商品 u, v に枝 $e_{u,v} \in E$ をはる．ここで σ はユーザが与える最小類似度である．その結果得られた一般無向グラフ $G = (V, E)$ を商品類似度グラフと呼ぶ．

2.2 NetSimile

NetSimile は，複数のネットワーク間の類似度を測定するために提案された手法である [Berlingerio 2012, Berlingerio 2013]．この手法は，1) 異なるサイズのネットワークに適用でき，2) 枝数に線形なスケーラビリティがあり，3) 節点や枝の対応関係がなくてもよい，といった特徴をもつ．

NetSimile は，1) 節点特徴量の抽出，2) グラフ特徴量の集約，3) グラフの類似度計算，の 3 つのステップから構成される．本研究では，ネットワークの類似度を測定する目的に NetSimile を利用するのではなく，最初の 2 つのステップで得られるグラフ特徴量を利用する．そして NetSimile によって列挙された特徴量から，後述の罰則付き回帰モデルによって目的変数に寄与する特徴量を抽出する．以下では，本研究に関連するステップ 1),2) について説明する (図 1 に概略図を示す)．

2.2.1 節点特徴量の抽出

商品類似度グラフ $G = (V, E)$ について，全ての節点 $v \in V$ について，以下に定義される 7 つの特徴量を計算する．これらの特徴量は節点に定義されるため，特に「節点特徴量」と呼ぶ

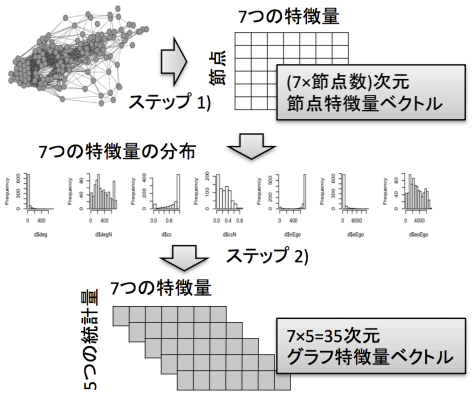

図 1: NetSimile の概略図

ことにする．以下の定義において，$N(v)$ は節点 v の近傍節点 (接続のある節点) の集合を表している．また $ego(v)$ は節点 v のエゴ・ネットワーク (「エゴネット」と略称する) を表しており，ここでは節点 v および節点 v から 1-hop で到達できる節点集合から誘導される G の部分グラフのことである．

1. 次数 $d_v = |N(v)|$：節点 v と接続のある節点数.
2. クラスタ係数 $c_v = \binom{d_v}{2}^{-1} |\{e_{u,w} | e_{u,w} \in E, u, w \in N(v)\}|$：近傍節点間の枝の数を近傍節点の 2 つの組合せで割ったもの.
3. 近傍平均次数 $\bar{d}_{N(v)} = \frac{1}{d_v} \sum_{u \in N(v)} d_u$：近傍節点の平均次数.
4. 近傍平均クラスタ係数 $\bar{c}_{N(v)} = \frac{1}{d_v} \sum_{u \in N(v)} c_u$：近傍節点の平均クラスタ係数.
5. エゴネット枝数 $eego_v$：エゴネット $ego(v)$ 内の枝の数.
6. エゴネット接続枝数 $eego_v^o$：エゴネット $ego(v)$ に接続される枝の数.
7. エゴネット近傍節点数 $nego_v$：エゴネット $ego(v)$ の近傍節点数.

2.2.2 グラフ特徴量の集約

前節の節点特徴量を抽出した段階で，節点 × 特徴量 行列が得られるが，次のステップでは，これらの特徴量を集約することでグラフ全体の特徴量 (「グラフ特徴量」と呼ぶことにする) を求める．グラフ特徴量は，7 つの節点特徴量それぞれについて，節点をサンプルと考えた場合の分布により定義される．ここには節点特徴量の分布の形状が，そのグラフを識別する「署名 (signature)」になるとの仮定がある．そして NetSimile では，分布の形状は，中央値，平均値，標準偏差，歪度，尖度の 5 つの統計量によって要約される．以上の集約により，7 つの節点特徴量について 5 つの統計量の 35 次元特徴量ベクトルが得られ，これをグラフ特徴量として用いる．

本実験では，一つの店舗について顧客別に商品類似度グラフを作成するため，全てのグラフ (顧客) において節点の意味は共通となる．そのため，前項での 7 つの節点特徴量も説明変数として利用することが可能である．ただし，購入実績のない商品の特徴量は 0 と定義した．

以上，ある店舗で扱われる商品数を M とすると $7 \times M$ 次元の節点特徴量ベクトルと 35 次元のグラフ特徴量ベクトルを全ての顧客について計算し，説明変数として構成した．

2.3 構造同値

一般無向グラフ $G = (V, E)$ の任意の 2 つの節点 $u, v \in V$ が構造同値性を持つとは，u と v が，G 上のその他の節点と完全に同じ構造を持つことを言う．構造同値性を持つ節点同士は，ネットワーク内で同じような役割にあると考えられる．しかしながら，完全に同じ構造を持つ節点ペアは現実のネットワークの中では稀であると言われている．そこで，節点間の類似度を定義し，ある閾値以上の類似度をもつ節点ペアは構造同値として扱うアプローチがとられることが一般的である．本実験においては，類似度の定義としてエゴネットの Jaccard 係数を用いる．

グラフ上での 2 つの節点 u, v のエゴネット Jaccard 係数 $jc(u, v)$ は，以下のとおり定義される．

$$jc(u, v) = \frac{|E_{ego(u)} \cap E_{ego(v)}|}{|E_{ego(u)} \cup E_{ego(v)}|} \tag{2}$$

ただし，$E_{ego(u)}$ はエゴネット $ego(u)$ の枝集合を表している．そしてユーザが与えた最小類似度 δ 以上の節点ペアを構造同値性のある節点として抽出する (本実験では $\delta = 0.7$ に固定)．そして抽出された節点ペア $u, v \in V$ のうち，接続のない節点ペア $\{(u, v) | e_{u,v} \notin E\}$ のみを選択する．これは，接続のない構造同値性を持つ商品ペアは競合関係にあるとの仮説があり，競合関係にある商品を購入するような顧客は，その店を日常的に利用していると考えられるからである．以上のようにして抽出された構造同値性をもつ商品ペアを 0-1 値をとる特徴量ベクトルとして構成した (「構造同値特徴量」と呼ぶ)．

2.4 モデル構築

以上により得られた節点特徴量，グラフ特徴量，そして構造同値特徴量を説明変数とし，顧客の来店回数の多寡を目的変数とした罰則付きロジスティック回帰モデルを構築する．

目的変数を $y \in \{0, 1\}$(0:来店回数が少ない顧客,1:多い顧客) とすると (多寡の定義は後述)，ロジスティック回帰モデルは式 (3) で表される．

$$\Pr(y = 1 | \mathbf{x}) = \frac{1}{(1 + \exp(-(\boldsymbol{\beta}^\top \mathbf{x} + \beta_0)))} \tag{3}$$

ここで，$\mathbf{x}$ は顧客の特徴量ベクトルである．$\boldsymbol{\beta} \in \mathbb{R}^p$, $\beta_0 \in \mathbb{R}$ は，それぞれ回帰係数ベクトルと定数項であり，式 (4) に示される罰則付き対数尤度を最小化することで推定される．

$$\begin{aligned} \mathrm{argmax}_{\boldsymbol{\beta}} \Bigg[\frac{1}{N} \sum_{i=1}^{N} \{ & y_i \log \Pr(y_i = 1 | x_i) \\ & + (1 - y_i) \log(1 - \Pr(y_i = 1 | x_i)) \} - \lambda P(\boldsymbol{\beta}) \Bigg] \end{aligned} \tag{4}$$

ここで，$J(\boldsymbol{\beta})$ は回帰係数に対する罰則項，$\lambda \in [0, \infty)$ は罰則項の重みである．λ は交差検証により未知データに対する推定平均二乗誤差を最小化するように調整する．罰則項 $J(\boldsymbol{\beta})$ として，本研究では elastic-net 罰則を利用した．elastic-net 罰則は，$J(\boldsymbol{\beta}) = (1 - \alpha)/2 ||\boldsymbol{\beta}||_2^2 + \alpha ||\boldsymbol{\beta}||_1$ で定義される．$||\boldsymbol{\beta}||_q$ は q-ノルムで $||\boldsymbol{\beta}||_q = (\sum_{i=1}^{p} \beta_i^q)^{1/q}$ である．$\alpha (0 \leq \alpha \leq 1)$ は，1-ノルムと 2-ノルムの重みであり，α の値が大きければ，

表 1: 特徴量別正解率

	横浜店		品川店	
特徴量	正解率	#p	正解率	#p
1-item 数量	0.856	115	0.900	211
2-item 数量	0.727	80	0.948	284
節点数	0.626	1	0.599	1
枝数	0.609	1	0.582	1
枝密度	0.636	1	0.595	1
グラフ特徴量	0.681	3	0.774	17
節点特徴量	0.670	50	0.799	112
構造同値特徴量	0.597	1	0.517	1

正解率は予測クラスと実クラスの一致割合，#p は選ばれた説明変数の数を示す．ランダムに予測した場合の正解率は品川店 0.52，横浜店 0.59 である．

係数が 0 と推定される変数が多くなり，逆に小さければ，全ての係数を全体的に小さくするように推定される．本研究では $\alpha = 0.8$ に固定して実験を行った．

3. 手法の適用

本研究で利用する購買履歴データは，あるスーパーマーケットの 2013 年 7 月から 2014 年 6 月までの 1 年間のデータで，品川と横浜の各 1 店舗を利用する．これらは共に同程度の売上と床面積を持っており，2 つの店舗を比較することで優良顧客の店舗別の違いも示す．また，分析対象とする顧客は，各店舗で 1 年間の来店回数が 50 回以上の顧客を対象に，90 回以上来店のある顧客を優良顧客，50 回以上 90 回未満を一般顧客と定義した．50 回以上の来店回数に限定した理由は，来店回数と購入商品数には正の相関関係 (相関係数 0.67) があり，一部の説明変数がモデル構築の際に影響を与えるため，来店回数の条件をいれることで説明変数と目的変数の依存性を排除している．

3.1 顧客別の類似度グラフの生成と特徴抽出

2.1 節の方法で，顧客毎に日別の購入商品群から類似度グラフを作成した．その際のパラメータ σ は, 横浜店，品川店ともに 0.005 と定めた．次に，各顧客から生成した類似度グラフに NetSimile を適用し，ノード毎に 7 次元の節点特徴量とグラフ全体の特徴量として 35 次元のグラフ特徴量を生成した．それ以外にも類似度グラフから，節点数，枝数，枝密度，構造同値特徴量を計算し説明変数として利用した．類似度グラフ以外の説明変数は，1 アイテムセット，2 アイテムセットの出現頻度を用いた．これらの説明変数と，優良顧客の判別を目的にして罰則付きロジステック回帰モデルを構築した．その際に利用した elastic-net 罰則の重みである α は，横浜店，品川店共に 0.8 と定めた．

表 1 は特徴量別の正解率を示している．正解率は予測したクラスと実際のクラスが一致した割合である．横浜店，品川店の両方で 1-item 数量の正解率が最も高いが，選択されている説明変数の数は 115 個，211 個とそれぞれ多く，全ての説明変数を解釈することは困難である．そこで， 両方の店舗で共通して優良顧客に寄与している変数を抜き出したものが表 2 である．両店舗に共通する商品で，売上金額ランキングの高い商品は，惣菜と加工肉で「のり弁当」,「コロッケ」,「ロースハム」は他の顧客と同様に優良顧客にも好まれる商品である．一方でパンカテゴリの中でもランキングの低い商品である「ずっしりリンゴデニッシュ」,「ナイススティック粒チョコ」は優良顧客が好む商品であり，売上ランキンが下位でもストックの必要な商品と判断できる．次に品川店にのみ出現し優良顧客に寄与する商品は,「プチチョコビスケット」や「チョコボールキャラメル」などの菓子と，タバコである．これは品川店というビジネス街にある店舗がその要因であると考えられる．また，横浜店の優良顧客に特徴的な変数は,「木綿豆腐」や「極小粒納豆」などの日配と,「そうめん専科」,「カルボナーラ」などの調味料が多く出現しており，賞味期限の短い日配や調味料を購入する主婦層が優良顧客として考えられる．

表 2: 横浜, 品川店に共通する変数

商品名	オッズ比	中分類	売上ランク
ずっしりリンゴデニッシュ	1.091	パン	215
ナイススティック粒チョコク	1.361	パン	380
彩りキッチンロースハム３Ｐ	1.011	加工肉	2
しらす干　並	1.003	塩干	10
のり弁当	1.004	惣菜	2
コロッケ	1.001	惣菜	3
しらたき	1.045	日配	46
こんにゃく　黒	1.023	日配	61
野菜かき揚げ	1.002	自家製惣菜	12
いなり 3ヶ入	1.038	自家製惣菜	96
にんじん	1.017	野菜	9
長ねぎ	1.011	野菜	13
アスパラ	1.018	野菜	16
キャベツ１／２切	1.008	野菜	66
さんま	1.033	鮮魚	8

品川店と横浜店の回帰モデルで共通して得られた 1-item 数量の結果．売上ランクは中分類別の売上金額のランキングを示している．

次に，表 3 に横浜店の優良顧客に寄与する 44 の節点特徴量を示す．優良顧客に寄与する変数の多くが，クラスタ係数や近傍接点の次数などであり，表 3 の商品と密に接続されている商品の関係が優良顧客の特徴として出現している．特に「コロネ練乳風味ホイップ」や「もち食感カスタードホイップ」のオッズ比が高く，このパンを中心にしたクラスタ係数の大きさが優良顧客の内在的な購買行動として現れている．

次に，グラフ特徴量について見てみると，驚くべきことに，節点特徴量との比較においてより少数の変数によって同等の精度を達成していることがわかる (表 1)．特に横浜店では，3 つのグラフ特徴量で節点特徴量のモデル精度をしのいでいる．この結果は，個々の商品についての購買行動を見なくても，商品全体の関係性に顧客の購買行動の特徴が現れ，それが優良顧客の内在的な購買行動として現れていると考えられる．また表 4 に各店における選ばれたグラフ特徴量一覧を示す．横浜店と品川店で共通して現れるグラフ特徴量は，近傍クラスタ係数平均だけであり，優良顧客に与える要因に関する一般理論の導出にはいたらないが，近傍密度が優良顧客の購買行動に影響を与えていることが伺える．

最後に構造同値特徴量について表 5 に示す．構造同値を捉えることで，商品の競合関係と把握することが可能であり，例えば,「塩さば」と「真あじ開き」は同じ日に購入されることは少なく，同様に「生姜焼き弁当」と「肉野菜炒め弁当」も同じ日に購入されることは少ない．このように，競合商品を把握できているが，構造同値を説明変数にしたときのモデルの正解率は低く，優良顧客の判別に構造同値はほとんど寄与していない．これは構造同値性を持つ商品が少ないことが原因と考えられる．

表 3: 横浜店の優良顧客に寄与する節点特徴量

商品	NetSimile	オッズ比	中分類名
もち食感カスタードホイ	cc	1.443	パン
塩あずき	cc	1.222	菓子
即席生みそ汁わかめ	cc	1.140	調味料
大陸肉ワンタンしょうゆ味	cc	1.120	惣菜
キャビンマイルド BOX	cc	1.031	タバコ
美山　焼き肉屋の味キムチ	cc	1.023	日配
おきあみかき揚げ	cc	1.013	自家製惣菜
出前一丁ゴマラー油５食	cc	1.009	麺類
コロネ　練乳風味ホイップ	ccN	1.597	パン
オランジーナ	ccN	1.452	飲料
すっきりチューハイレモン	ccN	1.271	洋酒
モチ食感粒餡小豆ホイップ	ccN	1.156	パン
イカフライ	ccN	1.085	自家製惣菜
くだものゼリーみか	ccN	1.080	菓子その他
なっちゃんオレンジ	ccN	1.028	飲料
もち食感カスタードホイ	degN	1.003	パン
アスパラ	degN	1.003	野菜
昆布ぽん酢	degN	1.003	調味料
生芋しらたき	degN	1.002	日配
タカトー　ちくわぶ	degN	1.002	練製品
焼肉のたれ　辛口	degN	1.002	加工肉
カキフライ	degN	1.002	塩干
キャビンマイルドＢＯＸ	degN	1.001	タバコ
新玉ねぎ	degN	1.001	野菜
麦穂の恵み	degN	1.001	パン
アジフライ	degN	1.001	惣菜
3 種のレタスサラダ	degN	1.000	野菜
骨なしフライドチキン	degN	1.000	加工肉
あぶり焼豚	degN	1.000	自家製惣菜
ふじっこ煮　ごま昆布	degN	1.000	惣菜
６Ｐチーズ	degN	1.000	酪農品
昭和　キャノーラ油	degN	1.000	調味料
長ねぎ	nEgo	1.001	野菜
天氷	nEgo	1.001	アイス
ふっくらおにぎり　さけ	nEgo	1.001	自家製惣菜
生姜焼き弁当	nEgo	1.000	自家製惣菜
そうめん専科	nEgo	1.000	調味料
チキン唐揚	nEgo	1.000	自家製惣菜

横浜店で優良顧客に寄与する節点特徴量．cc はクラスタ係数， ccN は近傍節点の平均クラスタ係数，degN は近傍節点の平均次数，nEgo は ego ネットワークに接続された節点数を示す．

表 4: 品川店と横浜店の優良顧客に寄与するグラフ特徴量

横浜店			品川店		
特徴量	統計量	回帰係数	特徴量	統計量	回帰係数
ccN	mean	-13.58	ccN	mean	-5.56
degN	ukurt	-0.0088	ccN	median	-2.84
deg	median	-0.0036	ccN	skew	-0.034
deg	ukurt	-0.0033			
deg	usd	0.0082			
deg	uskew	-0.013			
eEgo	mean	0.00019			
eEgo	ukurt	-0.00067			
eEgo	uskew	6.31			
eoEgo	median	-6.37			
eoEgo	ukurt	0.0068			
ccN	ukurt	0.044			
nEgo	median	0.00054			
ccN	usd	0.094			
cc	median	-0.10			
cc	ukurt	0.035			
cc	usd	-0.60			

表 5: 構造同値特徴量

横浜店
塩さば２枚 - 真あじ開き２枚
カーネーション - 新まるごとバナナ
生姜焼き弁当 - 肉野菜炒め弁当
いなり 3ヶ入 - よくばり４種おにぎり海鮮
よくばり４種おにぎり - 和風助六 (小)
プロッシモ完熟ホールトマ - 日清ディチェコペンネ
品川店
一平ちゃん夜店の焼そば - 赤いきつねうどん
一平ちゃん夜店の焼そば - 黒糖かりんとドーナツ
赤いきつねうどん - 黒糖かりんとドーナツ
アメリカ産豚バラ切り落し - 国内豚バラ薄切り
かつおたたき　刺身 - 中延園野菜炒め
りんご - カイワレ大根

4. おわりに

本研究は，来店回数から優良顧客を定義し，優良顧客の購買行動を表層的な関係から捉えるアイテム集合と，内在的な関係を捉える商品間ネットワーク構造を用いてモデル化した．表層的な関係性は，購買の直接的な影響を捉えることが可能であり，売上ランクは高くないが， 優良顧客に好まれる商品を明らかにした．一方で，商品間のネットワーク構造については，NetSimile を適用することでグラフ特徴量を算出し，ある商品から接続されている商品間の関係性が密になることが優良顧客の購買行動として明らかになった．また構造同値に関しては，構造同値性を持つ商品が少ないため，優良顧客を判別するほどのモデル化ができなかったため，この点は改善が必要である．

謝辞

JST-ERATO 湊離散構造処理系プロジェクトの中元政一氏，Stephane Cheung 氏，国立情報学研究所宇野 毅明氏からは有意義な情報と適切なコメントをいただいた．ここに感謝の意を表します．また，本研究の一部は，科学技術振興機構 CREST，及び ERATO 湊離散構造処理系プロジェクト，文部科学省の科研費若手研究 (B) 4730375 の研究助成を受けている．

参考文献

[Berlingerio 2012] M. Berlingerio, D. Koutra, T. Eliassi-Rad, and C. Faloutsos, “ NetSimile: A scalable approach to size-independent network similarity, ”CoRR, vol. abs/1209.2684, 2012.

[Berlingerio 2013] M. Berlingerio, D. Koutra, T. Eliassi-Rad, and C. Faloutsos. Network similarity via multiple social theories. In ASONAM, 2013.

[Sailer 1978] Lee Douglas Sailer, Structural equivalence: Meaning and definition, computation and application, Social networks, Volume 1, Issue 1, 19781979, Pages 7390

[清水 1996] 清水聰,「新しい消費者行動」, 千倉書房, 1996.

[清水 2004] 清水聰,「消費者視点の小売戦略」, 千倉書房, 2004.