2G1-1

買い物時間による顧客の分類と売場の訪問順序を考慮した購買モデルに関する研究

The Study of a Purchase Model for Shopping Time and Path Using LCM Sequence

中原孝信　　矢田勝俊
Takanobu NAKAHARA　Katsutoshi YADA

関西大学商学部
Faculty of Commerce, Kansai University

In this study, we use a dataset called radio-frequency identification (RFID) data to evaluate the customers' in-store movements. The RFID data, which is an extraordinary new dataset, show the position of a shopping cart through an RFID tag attached to the shopping cart. We can analyze customers' purchasing behavior and their in-store movement information using POS data combined with RFID data. The purpose of this study is to discover a promising shopping path that can distinguish customers' in-store movements by sequential pattern analysis using RFID data. These shopping paths are extracted using LCMseq, which is an efficient algorithm for enumerating all frequent sequence patterns. Finally, shopping paths are used in the purchasing model to generate the rules that expressed customers' in-store movements and purchasing characteristics. As a result, in this study, we came up with some useful suggestions for a more efficient in-store area management.

1. はじめに

顧客の小売店における購買行動を理解する試みとして、1990年代より、大規模な ID 付き POS データを蓄積し、そのデータを分析することで、有用な知見を得ようとする試みが積極的に行われている。しかしながら、これらのデータは購買行動の結果を示すだけであり、店内でどのような売場に訪れ、どれぐらい店内を歩き、どれぐらい考え、そして購買に至ったか、そのプロセスを明らかにしてくれるわけではない。マーケティング研究ではこのような購買プロセスはブラックボックスとして扱われてきた。

近年、新しい技術革新により、さまざまな場面で顧客の購買プロセスに関するデータが蓄積されるようになってきた。特に小売店では、RFID を用いて顧客の店内の動きを追跡したショッピングパスデータ [Sorensen 03] や、買物時の視線の動きを記録した eye tracking データ [Krugman 94]、[Fox 98] などがある。

ショッピングパスデータと ID 付き POS データを組み合わせることで、いつ，どの顧客が，どの商品をいくらで，どのような巡回経路と移動距離、そして時間をかけて購入したかを特定することができる。本研究では、この 2 つのデータを組み合わせて利用し、新しい購買行動モデルを提案することで、有用な知識の発見を目指す。具体的には、LCM シーケンス [Ohtani 08]、[Uno] と呼ばれる頻出部分シーケンスを高速に列挙できるアルゴリズムを用いて、高額購買顧客と低額購買顧客の巡回経路とその購買行動を識別できるパターンをデータから抽出する。その際、買物時間を基準に顧客をあらかじめグループ化しておき、同程度の買物時間を持つ顧客グループに対して LCM シーケンスを適用する。最終的に、抽出したパターンを用いて顧客グループ毎に判別モデルを構築し、高額購買顧客と低額購買顧客それぞれに特徴的な店舗内の購買行動を特定する。

連絡先: 中原孝信　関西大学商学部　吹田市山手町 3-3-35、
TEL:06-6368-1322、Fax:06-6368-1322、
nakapara@kansai-u.ac.jp

2. 既存研究と研究目的

ショッピングパスデータと ID 付き POS データを組み合わせて利用した研究として、Larson [Larson 05] らは、時間を基準（長、中、短）に顧客を 3 つにわけ、各グループに k-means を改良したクラスタリング手法を用いて、複数の顧客グループを発見し、それぞれのグループの典型的な買物経路、新たな仮説を導き出している。また、Hui [Hui 09] らは、オペレーションズ・リサーチにおける巡回セールスマン問題の枠組みを適用し、顧客の売場間移動と購入商品の関係を明らかにしている。

ショッピングパスに関するこれらの既存研究は、移動経路と購入行為の基本的な関係を明らかにしようとするものであり、顧客ごとの売場での訪問順序や買物時間などが購買に与える影響など、ショッピングパスに含まれる詳細な情報と購買の関係を十分に捉えることはできていない。これまで行って来た我々の研究では [Nakahara 10]、購入量の違う顧客を比較した際に、訪問する売場に大きな違いは見られず、売場の訪問順序や滞在時間などに重要な違いが存在することが分かっている。売場の担当者にとって、買物量を増加させる売場レイアウトの設計のために、そうした違いを理解することは重要である。また、データから得られた知見を実際の売場で活用するためには、より具体的な施策を打ち出すことができるように、特定の顧客群の差異を生み出す要因を明示化することが重要である。

そこで本研究では、ショッピングパスデータを対象にした分類問題への研究枠組みの提案と、明示的で有用なショッピングパスを発見し、顧客の買物行動に関する知見を得ることを目的とする。そのために、まず Larson [Larson 05] らの枠組みに従い、買物時間の観点から同程度の買物時間を持つ顧客をそれぞれ買物時間の「長」「中」「短」にグループ化する。次に、各グループの中で高額購買顧客と低額購買顧客の購買行動の違いを生み出す要因を明示化するために、訪問順序に着目した部分的なショッピングパスを抽出する。そして、売場の訪問順序と購買行動の関係を明らかにすることができるモデルを提案し、その有用性を検証する。

3. 訪問パターンの抽出と分類問題への枠組み

3.1 系列データベースへの変換

本研究は、売場の訪問順序に着目した特徴的な部分経路を発見することで、高額購買顧客と低額購買顧客を判別する要因を発見することを目的としている。そこで、実際のショッピングパスデータに含まれる X、Y 座標の位置から店内の各売場を特定し、売場をアイテムとして扱うことで、売場の訪問順序を考慮した系列データとして表現する。図 1 は、その変換を例示している。図 1 の a) は RFID から得られたデータであり、顧客 ID、日時、X、Y 座標の位置と、それから得られた売場などの項目が含まれている。そして、このデータに含まれる同一顧客の売場 ID を時系列順に並べることで、図 1 の b) のように系列データとして売場訪問順序が表現される。

顧客 ID	日付	時刻	X	Y	売場	売場 ID
0001	20090511	120342	95	531	入口	E
0001	20090511	120456	125	331	青果1	V1
0001	20090511	120458	155	271	青果2	V2
0001	20090511	120517	151	105	鮮魚2	F2
0001	20090511	120639	62	75	鮮魚1	F1
0001	20090511	120655	500	90	惣菜	G
0001	20090511	120658	500	96	一般食品6	B6
0002	20090511	120659	500	91	惣菜	G
0002	20090511	120737	499	142	一般食品6	B6
0002	20090511	120742	565	194	冷食	K
0002	20090511	120754	637	297	洋日配	I

a) ショッピングパスデータ

顧客 ID	売場訪問順序
0001	E,V1,V2,F2,F1,G,B6
0002	G,B6,K,I

b) 訪問売場順の系列データ

図 1: 訪問順序を示した系列データ

3.2 LCM シーケンスを利用した特徴的な訪問パターンの発見

LCM シーケンスは、系列データベースから、頻出系列パターンを高速に列挙できるアルゴリズムである。

ここで、任意のアルファベットを Σ、そして、Σ 上の有限系列全体を Σ^* と表す。系列パターンは任意の系列 $s = a_1 \cdots a_n \in \Sigma$ であり、$P = \Sigma^*$ で Σ 上の系列パターン全体の集合を表す。Σ 上の系列データベースは、系列の集合 $S = \{s_1, ..., s_m\}$ である。$\mid S \mid = m$ で S の要素数を表す。系列パターンが、ある系列の部分系列となるときに、その系列に出現するという。また、与えられた最小頻度値 $\sigma \geq 0$ 以上の数の系列に出現するときに頻出であるという。

頻出パターンをデータの分類に用いると、一方の集合に多く出現するが、他方の集合にはあまり出現しないパターンを 2 つの集合を特徴づけるパターンとして利用できる。本研究では、高額購買顧客と低額購買顧客をデータから定義し、2 つの顧客集合を特徴付ける部分系列をパターンとして抽出する。そこで、高額購買顧客集合の系列データに与える重みを w_h、低額購買顧客集合の系列データに与える重みを w_g とする。通常、頻出パターンを列挙する際には、各系列は全て等価であると見なし、重みは単一コストを用いて計算される場合が多い。しかし、LCM シーケンスでは、各系列に異なる重みを付与してパターンを抽出することが可能であり、その際、$\Sigma_{s \in Hc} w_h$ と $\Sigma_{s \in Gc} w_g$ の差が $minDiff$ 以上の系列パターンを抽出することが可能である。ここで、Hc と Gc は任意のパターン p が出現する高額顧客の系列データと低額顧客の系列データに対する部分系列集合をそれぞれ意味している。

このようにしてパターン抽出を行う場合には、集合の要素数が異なると問題が生じる。例えば、重みに単一コストを用いて、あるパターンが全ての系列に出現する場合を考える。そのパターンは、両方の集合に完全に含まれているため、一方の集合に特徴的なパターンではないが、要素数の多い集合に特徴的なパターンとして扱われてしまう。そこで、この問題を解決するために、$w_h = 1/\mid Hc \mid$、$w_g = 1/\mid Gc \mid$ という重みを導入する。1 を各集合の要素数で割ることで、$\Sigma_{s \in Hc} w_h$ と $\Sigma_{s \in Gc} w_g$ はそれぞれ 0 から 1 の範囲を取り、全ての系列に出現するパターンが存在しても、それらの差は 0 になる。したがって、要素数に依存することなく特徴的なパターンの抽出が可能となる。

3.3 時間、距離を考慮した系列パターン

訪問順序に加えて、各売場の滞在時間と移動距離をそれぞれ考慮した系列パターンを抽出する。これらを比較することによって、高額購買顧客と低額購買顧客を特徴付ける要因として順序、滞在時間、移動距離のどれが重要かを明らかにする。各売場の滞在時間と距離は、数値データであり、パターンを抽出する際には離散化が必要になる。そこで本研究では、図 2 のように、各顧客が訪れた売場の滞在時間を対象に件数が均等になるように 3 等分（5 秒以下、6〜18 秒、19 秒以上）に分割する。そして、訪れた売場の後に離散化した文字を付与することで時間を考慮した系列データとして表現する。各売場内で移動した距離も同様の方法で系列データとして扱う。

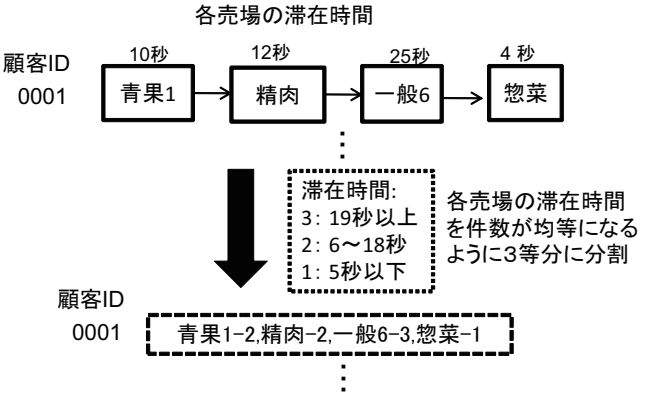

図 2: 訪問順序を示した系列データ

3.4 特徴的な訪問パターンによる判別モデルの構築

抽出したパターンを説明変数として利用し、最終的に高額購買顧客集合と低額購買顧客集合を識別する判別モデルを構築する。その際、訪問した売場の順序を示す系列データに抽出したパターンが含まれるなら 1、含まれないならば 0 とするダミー変数によって表す。また、パターン以外に利用する変数は、各売場の滞在時間比（各売場の滞在時間/1 回の買物総滞在時間）、各売場の移動距離比（各売場の移動距離/1 回の買物総移動距離）、滞在エリア種類数、滞在ポジション種類数、総滞在回数、来店間隔を用いる。そして、これらの説明変数を組み合わせて、以下のモデルを構築する。

1. 訪問順序の系列パターンだけのモデル
2. 訪問順序とその滞在時間を併せた系列パターンだけのモデル
3. 訪問順序とその売場内の移動距離を併せた系列パターンだけのモデル

表 1: 買物時間による分類と顧客の定義

買物時間	短（2 分以上 13 分未満）		中（13 分以上 20 分未満）		長（20 分以上）	
グループ	高額	低額	高額	低額	高額	低額
金額	2,312 円以上	2,311 円以下	3,184 円以上	3,183 円以下	3,184 円以上	3,183 円以下
人数	1,163 人	1,163 人	1,166 人	1,165 人	1,167 人	1,165 人

4. 訪問順序を考慮せずに抽出したパターンによるモデル化
5. 上記 1 から 4 で最も精度が高かったパターン変数とパターン以外の説明変数を用いた統合モデル

上記の 1 から 4 は、0,1 ダミー変数だけを扱うため数量化 II 類による判別分析、5 は、数値属性と 0,1 ダミー変数を両方扱うためロジスティック回帰によってモデルを構築する。

4. ショッピングパスデータへの適用

4.1 利用データと基礎分析

本研究で利用するショッピングパスデータは、日本国内の某スーパーマーケットチェーンの一店舗で得られたデータである。データの取得期間は、2009 年 5 月 11 日から 6 月 15 日までの約 1ヶ月間でショッピングカートに取り付けた RFID タグにより、約 7,000 人の顧客に関する巡回行動が特定できる。これらの顧客に関しては、ショッピングパスデータと紐付けられた ID 付き POS データの利用も可能である。ショッピングパスデータは、カートの利用者ごとに顧客が識別されるため、カートを利用した顧客のデータだけが蓄積されている。

図 3 は店舗内のレイアウトを示している。2 つの入り口と 25 の売場、そして、中央通路とレジから構成された売場となっている。対象店舗は、レイアウト図から確認できるように、外周に青果、鮮魚、精肉などの生鮮食品売場があり、内側には、菓子や一般食品などの売場から構成されている。これは、一般的なスーパーマーケットで見受けられるような売場の構成と同じレイアウトである。

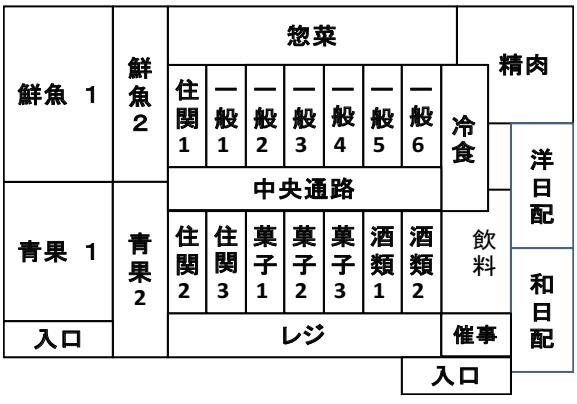

図 3: 店舗内のレイアウト

このデータに含まれる顧客は女性が多く、約 90%を占めている。また、顧客の年代は 30 代から 60 代が中心である。各売場によって、購買に関する特徴は異なっており、青果 1 は 25 個の売場の中で購買者数と訪問者数が最も多く、購入割合（購入者数/訪問者数）が 84%の売場であった。青果 1 に次いで、和日配、鮮魚 1、精肉の順に購買者数の高い売場であるが、訪問者数はそれぞれ、8 番目、11 番目、12 番目であり、これらの売場は、訪問者数はそれほど多くないが、訪れた顧客の多くがそこで商品を購入している売場である。一方、惣菜売場は訪問者数が 2 番目に高い売場であるが、購入割合は 47%で 8 番目の売場である。この売場は、訪問者数に対して購入者数が少ないことから、購入者数を増加させるために改善すべき売場である。このように、効率よく購買に結びついている売場やそうではない売場が存在している。

4.2 買物時間を考慮した顧客のグループ化と判別モデル

買物時間は、購買金額に何らかの影響を与えており、買物時間の長い顧客は、短い顧客よりも多くの購買金額を利用する傾向にある。したがって、高額購買顧客と低額購買顧客の特徴を抽出するに際して、購買時間によるバイアスを省く必要がある。ここでは、Larson [Larson 05] らの枠組みに従い、買物時間の類似した顧客を「短」「中」「長」の 3 グループに分け、各グループの中で高額購買顧客と低額購買顧客を定義し分析を行う。表 1 は、最終的に得られた分析対象となる顧客グループを示している。買物時間を人数が均等になるように 3 分割し、各時間グループの中で各顧客の総購買額を利用し、人数が均等になるように更に 2 つのグループに分けた。例えば、買物時間が短いグループは、2,312 円を基準に高額購買顧客と低額購買顧客に分類できそれぞれ 1,163 人が属している。これらの顧客集合が目的変数として利用される。説明変数に関しては、3.4 節で説明した通りであり、各買物時間グループの高額と低額購買顧客を特徴付けるパターンを LCM シーケンスによって抽出し、モデルを構築する。したがって、各買物時間グループの高額と低額購買顧客に対して、5 種類の判別モデルを構築するので、合計 15 種類の判別モデルを生成する。

4.3 判別分析の結果

表 2 は、買物時間グループが「長」で高額購買顧客と低額購買顧客の判別モデルの結果を示している。説明変数は、3.4 節の 1 から 5 の説明変数の組み合わせによるモデルを示している。各パターンを抽出するための売場は、図 3 のレイアウトに示す売場をまとめて同じ売場名で始まる売場は 1 つの売場として扱っている。例えば、鮮魚 1 と 2 は鮮魚として扱っており、同様に一般 1〜6 は一般として扱っている。これらの判別モデルは、SPSS18 の判別分析でステップワイズ法により計算を行った。順序を考慮せずに抽出したパターンだけを利用したモデルは、精度が最も低く 55.4%であり、1 から 4 のパターンだけを利用したモデルの中では、売場の訪問順序に時間を考慮したパターンを用いたモデルが 61.1%と最も高い精度であった。そのパターンにパターン以外のすべての変数を含めてモデルを構築した際の判別率は、63.2%になった。これらのことから、売場の訪問順序と滞在時間を考慮することが高額購買顧客と低額購買顧客を判別する要因として重要であることを示している。

表 3 は、買物時間グループが「長」を対象に最も精度の高かった「5. 訪問順序時間パターン＋全変数」を説明変数に利用した際の結果を示しており、いずれの変数も 5%有意であった。変数の先頭の ∗ は、その変数がパターン項目であること

表 2: 判別精度一覧

説明変数	判別精度
1. 訪問順序系列パターン	60.6%
2. 訪問順序時間パターン	61.1%
3. 訪問順序距離パターン	60.4%
4. 順序なしのパターン	55.4%
5. 訪問順序時間パターン＋全変数	63.2%

を示しており、モデルに多くのパターン変数が出現していることが確認できる。係数 B の値が正に大きい変数は、高額購買顧客を説明するために影響力のある要因であり、負の値が小さい変数は、低額購買顧客を説明するために影響力のある要因である。したがって、買物時間が長い顧客の中で、冷食や鮮魚の滞在時間の割合が大きいと、高額購買につながるが、入口の滞在時間が長くなると低額購買になるという傾向がある。またパターンを見ると、「青果-3, 鮮魚-3」と続くパターンは高額購買顧客を特徴付けるパターンとして出現しており、青果と鮮魚を連続して訪れ、いずれの滞在時間も長いほうが高額購買に繋がる傾向がある。紙幅の都合上省略させていただくが、買物時間の「中」「低」のグループに関しても、同じように解釈可能なモデルが生成できており、「中」に関しては、「訪問順序距離パターン＋全変数」が最も判別精度の良いモデルであり、「短」に関しては、買物時間グループが「長」と同じく「訪問順序時間パターン＋全変数」が最も高かった。

表 3: 判別精度一覧

順序＋時間と全変数	係数 B	有意確率
定数	-2.19	0
滞在時間比_冷食	4.033	0.001
滞在時間比_鮮魚	1.27	0.008
*青果-3_鮮魚-3_一般-3_一般-3_精肉-3	0.368	0.019
*青果-3_鮮魚-3_一般-1_和日配-3	0.308	0.005
*中央通路-1_一般-2	0.306	0.003
*青果-3_鮮魚-3_中央通路-1	0.198	0.046
滞在 Posi 種類数	0.082	0
総滞在回数	0.002	0.005
*レジ-3_青果-3	-0.346	0.007
*鮮魚-3_惣菜-2_青果-3	-0.365	0.014
*青果-3_一般-3_青果-3	-0.43	0.016
*惣菜-2_青果-2	-0.465	0
*鮮魚-3_一般-3_鮮魚-3	-0.554	0
滞在時間比_入口	-1.707	0.04

5. おわりに

本研究は、ショッピングパスデータと購買履歴データを用いて、買物時間を基準に 3 つのグループに顧客を分類し、それらのグループの中から LCM シーケンスにより高額購買顧客と低額購買顧客に特徴的な系列パターンの抽出を行った。また、それらのパターンを判別分析の説明変数として利用することで、各顧客集合を特徴付ける要因を抽出した。

本研究で示した動線データの適用例は、動線データに対する系列パターンの利用可能性を示したものであり、系列パターンによる特徴の抽出と、抽出したパターンを用いたモデルが有効であることを示した。また実際に得られたルールは、十分に解釈できるものであり、高額購買顧客と低額購買顧客の特徴を把握することができた。しかしながら、得られたルールのビジネス応用への可能性に関しては課題が残る。高額購買顧客の特徴が明らかになったが、それをどのようにしてビジネスに活かせば、最終的に売上の増加に結び付けることができるのか。という点を明確にすることができなかった。この点に関しては今後の課題としたい。

参考文献

[Sorensen 03] Sorensen, H., "The Science of Shopping", *Marketing Research*, 15, pp.30-35. 2003.

[Krugman 94] Krugman, D. M., Fox, R.J., Fletcher, J. E., Fischer, P. M. and Rojas. T. H., "Do adolescents attend to warnings in cigarette advertising? An eye-tracking approach." *Journal of Advertising Research*, 34(6),pp.39-52. 1994.

[Fox 98] Fox, R. J., Krugman, D. M., Fletcher, J. E., and Fischer. P. M., "Adolescents ' attention to beer and cigarette print ads and associated product warnings", *Journal of Advertising Research*, 27(3), pp.57-68. 1998.

[Larson 05] Larson, J. S. and E. T. Bradlow and P. S. Fader, "An exploratory look at supermarket shopping paths," *International Journal of Research in Marketing*. Volume 22, Issue 4, 2005, pp.395-414.

[Hui 09] Hui, S.K., Fader, P.S. and Bradlow, E.T., "Research Note -The traveling salesman goes shopping: The systematic deviations of grocery paths from TSP optimality", *Marketing Science*, 28, pp.566-572. 2009.

[Nakahara 10] Nakahara, T., T. Uno and K. Yada, "Extracting Promising Sequential Patterns from RFID Data Using the LCM Sequence", In proceedings of KES 2010, Lecture Notes in Computer Science, 2010, Volume 6278/2010, pp.244-253, 2010.

[Ohtani 08] 大谷英行, 喜田拓也, 宇野毅明, 有村博紀, 「極小出現区間を用いたエピソードマイニングの高速化」, 情報処理学会研究報告, 巻:2008 号:56 頁:113-120.

[Uno] http://research.nii.ac.jp/ uno/code/LCMseq.htm