2A3-1

LCM シークエンスを用いた顧客動線データからの特徴抽出

Pattern Mining in Shopping Path Data Using the LCM Sequence

中原孝信[*1] 宇野毅明[*2] 矢田勝俊[*3]
Takanobu NAKAHARA Takeaki UNO Katsutoshi YADA

[*1][*3]関西大学商学部 Faculty of Commerce, Kansai University
[*2]国立情報学研究所 National Institute of Informatics

Recently, supermarkets use radiofrequency identification (RFID) tags attached to shopping carts to track customers 'in-store movements and to collect data on their paths. Path data recorded customers' movements in a spatial configuration and contain valuable information for marketing, e.g. shopping trip time and distance, as well as number of shelf visits. We analyze customers' purchase behavior and their in-store movements not only by using path data, but also combining it with pos data. However, the volume of path data is very large, since x and y coordinates expressing position of a cart are updated every second. Therefore, we have to use an efficient algorithm in order to handle these data. In this paper, we apply LCM sequence to shopping path data to extract promising sequential patterns with the purpose to express prime customers' in-store movements compared with general customers. LCM sequence is an efficient algorithm for enumerating all frequent sequence patterns. Finally, we construct decision tree model using extracted patterns in order to discover prime customers' in-store movements.

1. はじめに

小売業では顧客の購買行動を理解する試みとして，ID 付き POS データなど顧客の購買履歴データを対象にした研究が多く行われてきた．近年では更に，RFID と呼ばれる認証技術を利用したタグをショッピングカートに取り付けることで，顧客の店舗内における巡回行動を顧客動線データとして蓄積する試みが行われている [Larson 05],[Yada 09]．本研究は，ID 付き POS データと店舗内の顧客動線データを組み合わせて利用しており，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によって購入したかを特定することが可能となる．しかし，顧客動線データは，数秒ごとにカートの位置が座標として更新されるため，蓄積されるデータ量は非常に膨大なものとなる．したがって，それら大量のデータを高速に扱える必要がある．

本研究では，LCM シークエンス [Ohtani 08],[Uno] と呼ばれる頻出部分シークエンスを高速に列挙できるアルゴリズムを用いて，高額購買顧客とそれ以外の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抽出する．シークエンスを解析することで，どのような順番で物を買い，どのように迷いが生じるのかなど，これまで得ることのできなかった買い回り行動を明らかにすることが可能となる．最終的に，抽出したパターンを用いて決定木モデルを構築し，高額購買顧客に特徴的な買い回り行動を特定する．そして，抽出したパターンを説明変数に利用することで，モデルの精度が改善できることを示す．

2. 分析対象データと基礎分析

本研究で利用する動線データは，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，2008 年 9 月末から 10 月末までの一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 1,000 人の顧客に関する巡回行動が利用できる．動線データは，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点に注意されたい．一方で，顧客購買履歴データは，同一店舗の 2008 年 7 月から 10 月末までの 4ヶ月間のデータで，約 25,000 人の購買データが利用可能である．動線データに含まれる顧客の中には，購買履歴データと併合できない顧客が存在したため，最終的にこれら 2 つのデータが利用できる 625 人の顧客を分析対象顧客とした．したがって，それらの顧客に関しては，動線データと購買履歴データの両方を利用することが可能である．

図 1 は店舗内のレイアウトを示しており，店内を 16 のエリアに分類したものである．対象店舗は，レイアウト図から確認できるように，左上にカート置き場があり，野菜売場（V）の横に，豆腐，納豆などの日配品売場（G），その横に肉売場（M），そして魚売場（F），惣菜売場（Z）と続き，レジ（R）へと到達できるような売場構成になっている．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．

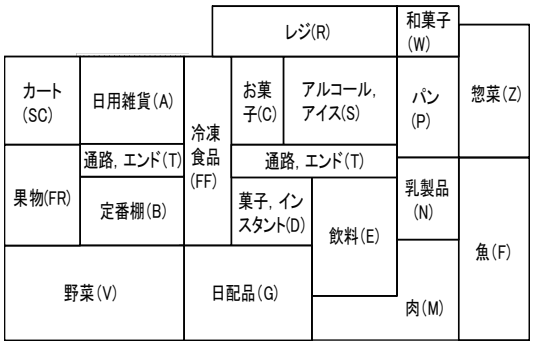

図 1: 店舗内のレイアウト

基礎分析の結果から，性別に関しては男女間の偏りは大きく約 90 %を女性が占めている．また年代別男女構成人数は，30 代から 60 代が中心である．動線データに含まれる顧客に関しては，30 代に比べて 60 代が若干多くなっている．これはカー

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

トの利用者から得られたデータであるため，若者の利用者に比べて，年配者が増加する傾向にあると考えられる．また同様の傾向として，一回の買い物で利用する平均購買金額なども若干動線データに含まれる顧客の購買金額が高くなる傾向があった．しかしながらデータとしては，顧客購買履歴データに含まれる顧客と比較して，それほど大きな違いは見られないことから，本分析のサンプルとして店舗内顧客動線データに含まれる顧客は十分に利用可能である．

次に店舗内の売場に関しては，売場毎にいくつかの特徴がある．野菜売場は訪問者数と購入割合（=購入者数/訪問者数）が 16 売場の中で最も高く，約 78%の購入割合である．また，豆腐や納豆などを扱う日配品売場，そして肉売場が野菜売場に続いて訪問者数と購入割合の高い売場である．これら 3 つの売場は，買い物客の多くが訪れ，そして訪れた多くの顧客が購入している売場である．惣菜売場は，訪問者数は 5 番目に高いが，購入割合は約 38%で 10 番目の売場であり，訪問者数に対して購入にあまり結びついていない売場であるといえる．このような売場は，たくさんの顧客が訪れていることから，レジまでの経路として利用されている可能性があり，商品の品揃えなどを改善することによって，更なる売上増加の可能性を秘めている．一方，パン売場は訪問者数が 13 番目であるが，購入割合が約 56%で 4 番目に高く，売場に訪れる人は相対的に少ないが，訪れた人たちの多くが購入している売場である．このような売場へ訪れる顧客は，最初から購買目的を持って訪れていると考えられる．

図 2 は，店舗内の巡回経路を有向グラフで示したものである．ノードは各売場を示しており，各売場を結ぶパスは売場間の移動を示している．そしてノードの大きさと各売場を結ぶパスの太さは，相対的な売場の訪問頻度と売場間の移動頻度を表している．ここで，頻度は立ち止まった行為だけを対象にしており，図示している経路は，最大移動頻度に対して 5%以上の経路だけを示している．相対的に移動の多い経路としては，カート置き場から，野菜，日配品，肉，魚類，惣菜，和菓子，そしてレジに至る経路であり，カート置き場から左回りに外周を移動する顧客が多いことが確認できる．また内周への経路としては，野菜売場，定番棚，そして通路を通り冷凍食品へ至る経路などが相対的に多い．しかし，内周は外周に比べて移動頻度の少ない経路が存在しており，そのような経路を活性化することが売上を増加させるためには重要である．

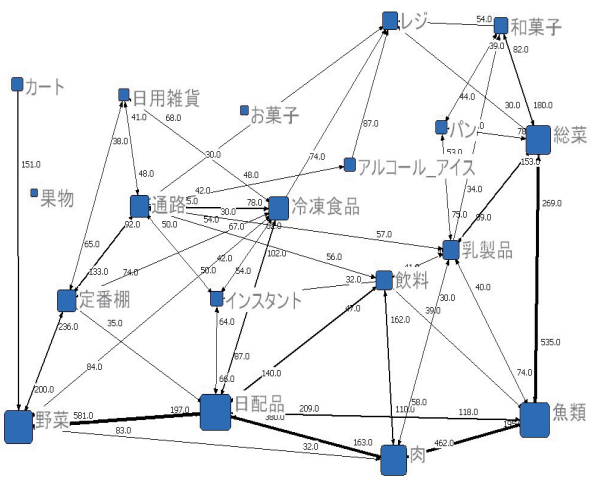

図 2: 店舗内の巡回経路

これら基礎的な分析から，各種売場には，商品を購入する目的を最初からもって訪れる売場と，商品を見てから購入を検討する売場，そしてレジに行くまでの巡回経路として訪れている売場などがある．レジに行くまでの巡回経路として訪れる売場は，訪問者数が多く，それらの売場で購買に結びつけることが売上増加のためには重要となる．

そこで本研究では，顧客を購買金額の多い顧客（高額顧客）とそれ以外の顧客（一般顧客）に分類し，それら 2 つの集合における店舗内の購買行動の違いを巡回経路の観点から識別して，高額顧客に特徴的な買い回り行動を LCM シークエンスを用いて発見する．

3. LCM シークエンスを利用した特徴的な巡回経路の発見

3.1 分析対象となる顧客集合の決定

625 人の顧客を高額顧客と一般顧客に分類する際に，動線データを取得する前の 3 カ月間の購買履歴データを利用して，全顧客の購買金額によるデシル分析を行った．表 1 は，全顧客の人数を均等に 10 分割したときに得られた各ランクにおける購買金額の範囲を示している．各ランクの顧客人数は，約 2,220 人ずつに分類されている．表の対象顧客数は，動線データに含まれる顧客が各ランクに属している人数を示しており，625 人の中で 607 人はいずれかのランクに存在していた．しかし，残り 18 人の顧客は動線データ取得期間にのみ来店していた顧客であったため分析対象顧客から省いた．人数と購買金額の関係から，高額顧客はランク 1 に属する 285 人とし，残りの 322 人を一般顧客として分析を行う．ランク 1 に属する高額顧客は，平均すると一ヶ月に約 1 万 5 千円以上利用している顧客である．以下では，動線データ取得期間のデータだけを利用して，高額顧客と一般顧客を比較し，各顧客集合に特徴的な巡回行動を発見する．

表 1: 購買金額デシル

ランク	購買金額範囲	対象顧客数
1	46,974 以上	285
2	26,345〜46,973 以下	116
3	16,245〜26,344 以下	73
4	10,245〜16,244 以下	52
5	6,552〜10,244 以下	32
6	4,155〜6,551 以下	20
7	2,612〜4,154 以下	12
8	1,573〜2,611 以下	6
9	807〜1,572 以下	6
10	806 以下	5

3.2 LCM シークエンスの適用

LCM シークエンスは，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．また，LCM シークエンスでは高速性に加えて，各系列に対して正負の重みを付与できることや，指定した窓幅に出現する頻出系列パターンだけが抽出できるなど幅広い適用が可能である．

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．$|S| = m$ 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．このとき，2 つの集合に含まれるパターン p の出現頻度の差がある閾値以上のパターンをコントラストパターン [Bay 99] と呼び，パターン p の出現頻度の割合がある閾値以上のパターンをエマージングパターン [Dong 99] と呼ぶ．

本研究では，動線データから頻出系列パターンを抽出するに際して，カート置き場からレジに至るまでに訪れた売場を時系列に並べることで，顧客の売場に対する巡回行動を系列データベース S として扱う．分析対象顧客は 607 人であり，$|S| = 607$ になる．次に，高額顧客集合の系列データに与える重みを w_h，一般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 シークエンスでは，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一般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．したがって，これはコントラストパターンを抽出することと等しい．

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，$w_h = 1/|Hc|$，$w_g = 1/|Gc|$ 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．本研究では，更に LCM シークエンスのもう 1 つの機能である窓幅（$1 \leq win \leq n$）を利用することで，顧客がある売場を訪れてから win 以内の売場の巡回行動に限定して 2 つの顧客集合に特徴的な系列パターンの抽出を試みる．

4. 計算結果

4.1 系列パターンの抽出

図 3 は LCM シークエンスにより $minDiff = 0.05$ 以上で抽出された系列パターンを示している．また，パターン抽出する際に，アイテムを付け足しても頻出度が変わらないパターンは，他のパターンに含まれるため冗長と判断して除外している．各点は 1 つの系列パターンであり，点の色はそれぞれ窓幅 win の値を表している．「その他」は，窓幅を n にした場合に抽出されたパターンであり，窓幅の制約をなくした状態と等しい．抽出された系列パターンの数は，全部で 9,622 個であった．図の横軸は一般顧客の重み合計値，縦軸は高額顧客の重み合計値をそれぞれ表している．原点を通る 45 度線を引いた際に，その線よりも上部は高額顧客集合に特徴的な系列パターンであり，下部は一般顧客集合に特徴的な系列パターンである．図から高額顧客に特徴的な系列パターンのほうが多く，高額顧客は一般顧客に比べてより多様な購買行動をしていることが考えられる．顧客集合の特徴を系列パターンで判別するためには，一方の顧客集合の重み合計値が高く，他方が低い系列パターンが好ましい．例えば，「定番棚_日配品 (購)」というパターンは窓幅が 5 の時のパターンであり，定番棚に滞在してから 5 つの売場内で日配品売場に滞在し，そこで商品を購入したという意味のパターンである．このパターンは，一般顧客の重みが大きく，一般顧客に特徴的なパターンの 1 つである．一方で，「野菜 (購)_魚」というパターンは，窓幅が 10 で，野菜売場で購入してから 10 売場内に魚売場に滞在するというパターンを示している．これは，高額顧客の重み合計値が 0.7 で，一般顧客が 0.57 なので，約 0.13 の差があるパターンで比較的判別力のある系列パターンである．例示したパターンは，滞在した売場と購買した売場の両方からなるパターンであり，POS データなどの購買履歴データだけでは得ることができないパターンである．これらは動線データを利用して初めて得られたパターンである．

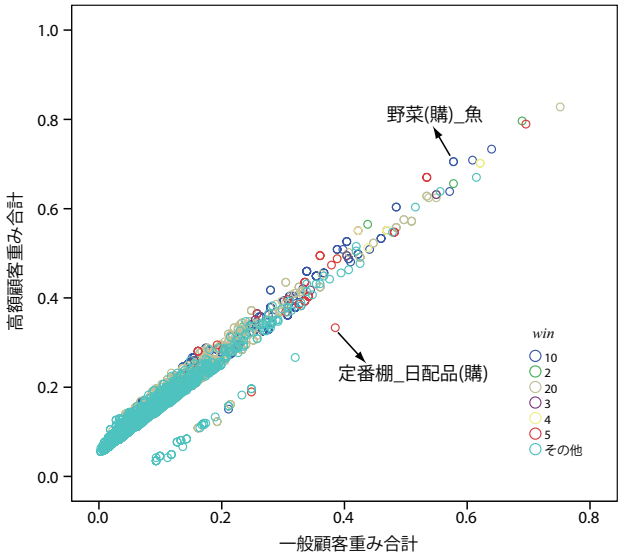

図 3: 抽出された系列パターン

4.2 系列パターンを用いた決定木分析

LCM シークエンスで抽出した系列パターンを用いて，決定木分析を行い，高額顧客と一般顧客の特徴を明らかにする．そこで，抽出された約 1 万のパターンの中から $minDiff$ の値を大きくし，高額顧客に特徴的な系列パターンとして $minDiff = 0.1$ 以上である 137 個のパターンを抽出した．一方で，一般顧客に特徴的な系列パターンは，最大で $minDiff = 0.07$ であり，パターン数が全部で 36 個と少なかったため，上述の系列パターンを全て利用した．したがって，説明変数として 173 個の系列パターンを 0,1 のダミー変数として利用し，それ以外にエリア別滞在回数，エリア別滞在時間割合，エリア別購買金額，購買エリアの種類数，滞在エリアの種類数，購買と滞在割合，合計買物時間 (分)，総滞在回数，総購買金額，総購買数量，平均購買金額，そして平均購買数量の合計 13 種類の説明変数を用いて，高額顧客と一般顧客の判別を目的に決定木分析を行う．

図 4 は決定木により生成されたモデルを示している．モデルの精度は，10 回の交差検証を行った結果 62.43%であった．系列パターンを含めずにモデルを生成した場合は，同様の検証

結果で 57.83%の精度であり，抽出した系列パターンを含めることで精度が向上した．

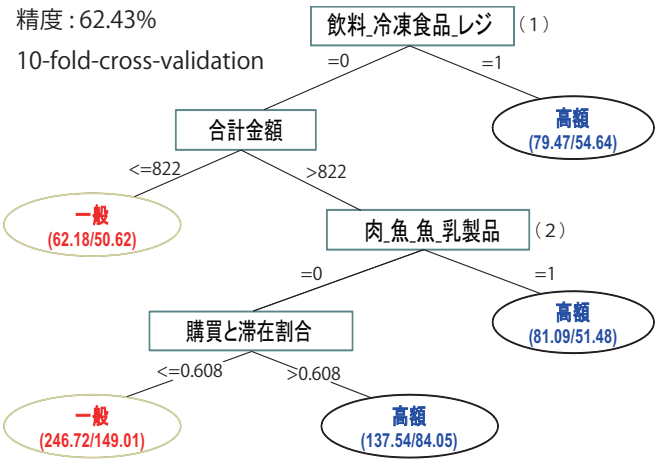

図 4: 高額顧客と一般顧客の決定木モデル

(1) のパターンは，飲料，冷凍食品，そしてレジの順に売場を滞在していることを示すパターンであり，このパターンは高額顧客を識別する特徴的なパターンである．各葉に記述している数値は，葉に到達するまでのルールに当てはまる件数と正判別数を示しており，このルールの判別率は 68.7%と比較的高い．このパターンを解釈すると，レジに行く前に飲料売場，冷凍食品売場の順に滞在しており，これは温度に敏感な商品を購入する際に見受けられる一般的な買い物傾向であり，このような行動をきっちり行う顧客は，高額顧客に多いと考えられる．

次に（2）のパターンは，肉売場，魚売場を 2 回，そして乳製品売場への滞在を表したパターンである．一回の買い物合計金額がある程度高く (822 円より高く)，このパターンを持つ場合は高額顧客に特徴的な購買傾向である．スーパーマーケットで主要な売上を占める生鮮食品売場を複数回に渡り滞在していることが重要であり，このパターンを持つ顧客は，この店舗にロイヤルティを持っていることが考えられる．

最後に最も下の分岐である「購買と滞在割合」は，購買売場数に対する滞在売場数の割合を表しており，この値が高いと立ち止まった売場では購買する傾向が強く，逆に低いと立ち止まっても購買には結びつかず，購買への迷いや，ためらいの傾向を表す．

これらルールの意味を踏まえて決定木を解釈すると，高額顧客の特徴を次のようにまとめることができる．

- 飲料，冷凍食品，レジの順で滞在する一般的な購買行動をきっちり行う顧客
- 購買金額がある程度高く，生鮮食品売場で複数回滞在する店舗にロイヤルティを持った顧客
- 購買金額がある程度高く，立ち止まると商品を購入しやすい顧客

逆に一般顧客の特徴は，

- 購買金額の低い顧客
- 購買金額はある程度高いが，立ち止まった売場で購買をためらう顧客

このように顧客の特徴を明確にすることで，高額顧客と一般顧客の店舗内の巡回行動，そして購買傾向を把握することが可能となる．このような特徴は，動線データを用いて初めて明らかになった点であり，購買履歴データとともに利用することで，これまで把握することが困難であった購買以外の行動を明らかにすることができる．

5. おわりに

本研究は，店舗内の巡回行動を蓄積した動線データと購買履歴データを用いて，LCM シークエンスにより高額顧客と一般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決定木分析の説明変数として利用することで，パターンを利用しない場合に比べて，判別精度を約 5%向上させることができた．

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分類問題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，高額顧客と一般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，最終的に売上の向上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

[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.

[Yada 09] Yada, K., ”String analysis technique for shopping path in a supermarket,” *Journal of Intelligent Information Systems*. 2009.

[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

[Bay 99] Bay, S. D. and M. J. Pazzani, “ Detecting change in categorical data: Mining contrast sets, ” *In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, 1999, pp.302-306.

[Dong 99] Dong, G. and J. Li, “ Efficient mining of emerging patterns: Discovering trends and differences, ” *In Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999, pp.43-52.