統計数理(2008)
第 56 巻 第 2 号 199–213

特集「データマイニングと統計数理」
[研究ノート]

スーパーマーケットにおける顧客動線分析と文字列解析

矢田 勝俊†

(受付 2008 年 1 月 4 日；改訂 2008 年 9 月 16 日)

要　　旨

本研究の目的は，顧客動線分析へ文字列表現を導入することで，文字列解析技術の適用可能性を検討し，その技術的な課題を明らかにすることである．実験では RFID 技術を用いて収集した小売業における顧客動線データに文字列解析技術を適用し，有用な知見の抽出を試みた．顧客動線データにおいて，我々は顧客の売場への立ち寄りに焦点を当て，売場訪問パターン文字列を生成し，買上数量の多い顧客の訪問パターンの特徴を明らかにした．そして実験を通して，顧客動線分析における文字列解析技術の課題などを明らかにすることができた．本論文では，顧客動線分析における文字列表現の適用可能性を指摘することができた．

キーワード： スーパーマーケット，マーケティング，RFID，文字列解析，EBONSAI.

1. はじめに

RFID は急速な技術進歩と低価格化が進むことによって，多様なビジネスに用いられるようになった．近年，その用途はマーケティングなどの小売業にまで拡大している．経済産業省における電子タグ実証実験においては，『日本版フューチャーストア・プロジェクト(経済産業省, 2005)』として，RFID のついたカートを用いて顧客の店内行動を把握し，店頭の販売活動ならびに顧客行動に関する情報を収集している．このプロジェクトでは，顧客の移動経路に関するデータが電子的に蓄積されており，従来ではブラックボックスであった店内の顧客の購買行動に関する詳細な情報を獲得している．こうした動きは日本だけでなく，欧米各国においても見られ，店内の顧客行動に関する RFID を用いた詳細な情報収集が注目を集めている．RFID データは，ストリームデータ(またはデータストリーム)と呼ばれ，データサイズが大きく，構造化されていないことが多いため，従来の研究で主に扱われていたような表形式データを対象にした手法を直接，適用することは難しい．流通分野や通信分野ではこうしたデータから有用な知見を得たいというニーズは高く，またデータマイニングの重要な応用領域としても，多くの研究者の関心を集めている(有村, 2005)．

従来，小売業における消費者行動の理解のために，POS データのような顧客の購買履歴データが用いられてきた．これらを利用すれば，どの顧客がいつ何をどれだけ購入したかが分かり，その購買行動を詳細に分析することが可能になる．例えばマーケティング分野では，Guadagni and Little (1983) や Gupta (1988) のように消費者の購買行動モデルが数多く提案されている．近年では，こうした膨大なデータを扱うために，多くの企業でデータマイニングが導入され

† 関西大学 商学部：〒564–8680 大阪府吹田市山手 3–3–35

(Hamuro et al., 1998; 矢田, 2004a)，店内の販売促進活動やブランド力の強化に役立てられている．しかしながら，顧客の購買履歴データはその顧客の購買結果が記録されているものの，彼女たちがどのように店内を移動しそれらを購入したのかを知ることはできない．既存研究において，店内における顧客の移動経路はブラックボックスとして扱われ，主にその結果である購買データだけが分析対象とされてきた．

近年の RFID 技術の進展はこうした状況を一変させている．特に RFID 技術のマーケティング応用研究の中で，最も注目されているのは，顧客または顧客のカートに RFID を付与し，店内の購買行動，移動経路を解析する顧客動線分析(Sorensen, 2003)である．従来のマーケティング研究のように購入商品という結果からではなく，顧客の店内での移動経路から新しい店頭販売活動の知見を得ようとするものである．こうした顧客動線に関する客観的なデータに基づいた研究は実は極めて少ない．なぜなら従来，そのようなデータを収集することが極めて困難であったためである．したがって RFID を用いて得られる顧客動線データは，マーケティング研究における新しい研究フロンティアの源泉を提供するものと思われる．

このような RFID を用いた顧客動線分析の既存研究として，Larson et al. (2005)の研究がある．彼らは顧客が利用するカートに RFID を付与し，店内の顧客の移動軌跡に関するデータを解析した．彼らは k-means を改良したクラスタリング手法を用い，複数の顧客グループを発見し，それらを詳細に検討することで，様々な仮説を提示している．しかしながら，顧客動線データからの特徴抽出や分類問題に焦点を置いた研究，適用事例はいまだ存在しない．小売業の現場では，所与のマーケティング戦略から導き出されたターゲット顧客について，その特徴把握や購買行動の理解に関するニーズは大きい．したがって，クラスタリング問題だけでなく，分類問題や特徴抽出に焦点を置いたアプリケーション研究は重要なビジネス・インプリケーションがあるものと思われる．しかしながら，マーケティング研究において，データに基づいた顧客動線分析は緒に就いたばかりで，有用な枠組みや知識表現は十分に議論されていない．

そこで我々は本論文において，顧客動線情報を含むストリームデータに対して文字列の知識表現を導入し，ビジネス分野での文字列解析のアプリケーションである EBONSAI (Hamuro et al., 2002; Yada et al., 2007)の適用を提案する．つまりストリームデータから顧客の店内の移動経路に関する情報を抽出し，それらを文字列として表現することで，既存の文字列解析手法を用いてルール抽出を行おうと考えた．本研究の目的は，顧客動線分析へ文字列表現を導入することで，文字列解析技術の適用可能性を検討し，その技術的な課題を明らかにすることである．

2. 顧客動線分析と文字列解析

2.1 顧客動線分析と文字列解析

顧客動線分析とは，店舗内を顧客が移動した経路を分析することによって，店舗レイアウトの設計や店内の販売促進計画の効率化を行おうとする店舗マネジメント手法の 1 つである．図 1 は実際の顧客動線を店舗レイアウト上に表現したものである．顧客が移動した経路とその方向は矢印付のリンクで表現されている．また，顧客が立ち止まった場所はノードで表され，赤のノードは何らかの商品購入を意味する．図から分かるように，顧客は極めて複雑な経路を移動し，買い物を行っている．

買物行動で特に重要なのが売場への立ち寄り率，つまり顧客がある売場を通った際に立ち止まったかどうかである．顧客はあるものを購入するためにカートを止め，その商品を手に取り，カートへ入れる動作が必要になる．それが売場への立ち寄りとしてデータ上，表現される．顧客の中には，売場で立ち止まったが，実際には商品を購入しなかった場合も存在するだろう．そのような顧客の購入の有無は動線データと販売データを比較すれば容易に判別できる．こうした情報は店頭マーチャンダイジングを行う企業にとって，極めて重要な情報となり得る．し

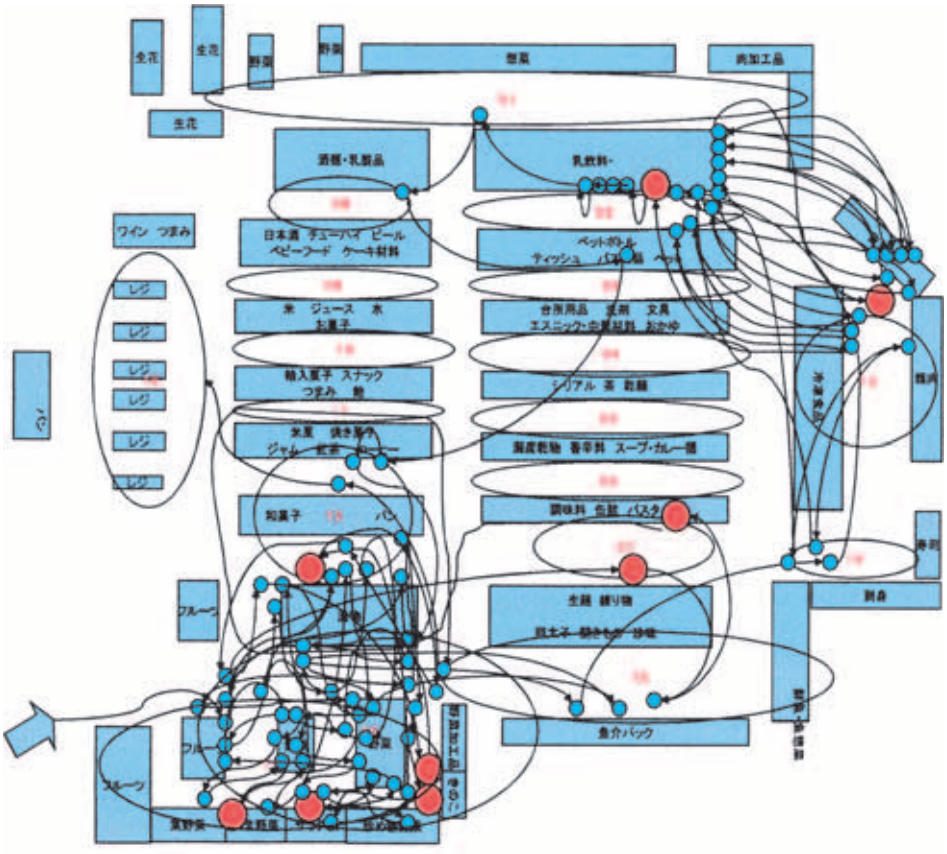

図 1. 顧客動線データの事例．赤丸は商品購入，青丸は立ち止まった場所を表す．

たがって，我々は本稿において，顧客動線データ内の売場への立ち寄りに焦点をあて，顧客の店内移動経路の特徴を抽出する．

RFID から得られるストリームデータはそのままでは処理が難しいため，何らかの加工の工夫が必要である．そのため本論文では，顧客動線データを分析するための知識表現として文字列を用いる．図 2 を用いて，その変換プロセスを説明しよう．図 2 の a) は RFID から得られる元データである．顧客識別情報や時間，X，Y 方向への加速度，カートの状態，受信した RFID のタグ No など様々な項目が含まれている．この元データを，売場レイアウトと RFID タグを関連付け，フロア ID を付与したレイアウト・マッピングテーブル b) を用いて変換する．RFID の各レコードは，各フロアにユニークな 1 文字に変換される．その際，顧客がその売場に立ち止まっているデータだけに絞り込む．最後に訪問順番を基準にフロア ID を結合することによって，文字列パターン d) が得られる．例えば，図 2 の顧客 Nancy の訪問パターンは b) のマッピングテーブルを用いると “AACFM” と表現される．

2.2 研究の目的

本研究の目的は，顧客の店内の購買行動に関するストリームデータを用いて行う顧客動線分析に文字列表現を導入し，その適用可能性を示すことである．本研究では，実際の RFID デー

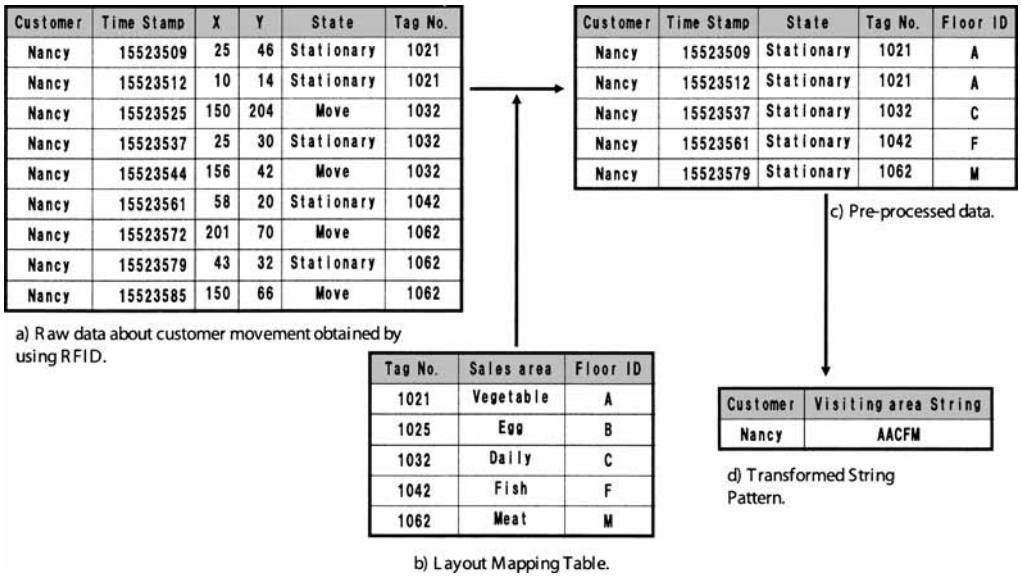

図 2. RFID データと訪問パターン文字列.

タから売場訪問パターン文字列を生成し，既存の EBONSAI システムを適用することで有用な知見を抽出する知識発見システムの構築を試みた．EBONSAI (Hamuro et al., 1998; Yada et al., 2007) とは，ゲノム解析などに用いられた文字列解析手法 BONSAI システム (Arikawa et al., 1993; Shimozono et al., 1994; Hirao et al., 2003) をビジネス用に改良した時系列解析システムである．従来，EBONSAI は販売データや WEB ログデータなどの時系列解析に用いられていたが，RFID から発生するようなストリームデータへの適用は行われていない．本論文では，こうしたストリームデータへの文字列解析手法の適用を試み，顧客動線分析における文字列表現の有効性を示し，従来の文字列解析技術 EBONSAI の技術的な課題を明らかにすることで，マーケティング分野におけるストリームデータの可能性を明らかにしたい．

2.3 EBONSAI

EBONSAI は分子生物学の分野で開発された文字列解析手法 BONSAI システム (Arikawa et al., 1993; Shimozono et al., 1994; Hirao et al., 2003) を改良したもので，文字列で表現された正負事例が与えられたとき，その部分文字列または部分シーケンスを用いて高精度に分類する決定木を生成するシステムである．まず EBONSAI のコアとなる BONSAI のアルゴリズムについて，簡単に説明しよう．

正事例集合を P，負事例集合を N，そして $|P||N|$ をそれぞれ P と N における事例数とする．部分文字列 α が与えられたとき，p_T，n_T を P と N それぞれの α が含まれる事例数とし，そして p_F，n_F を P，N で α が含まれない事例数を指すものとしよう．情報量関数 $ENT(x,y)$ は以下のように定義でき，

$$ENT(x,y)=\begin{cases}0 & x=0 \text{ or } y=0\\ -x\log x-y\log y & x,y\neq 0\end{cases} \tag{2.1}$$

部分文字列として α をデータが含んでいるかどうかで，元のデータを 2 つの部分集合 (subsets)

に分類した後の情報量を次のように定義する.

$$\frac{p_T+n_T}{|P|+|N|}ENT\left(\frac{p_T}{p_T+n_T},\frac{n_T}{p_T+n_T}\right)+\frac{p_F+n_F}{|P|+|N|}ENT\left(\frac{p_F}{p_F+n_F},\frac{n_F}{p_F+n_F}\right) \tag{2.2}$$

この値の最も小さな α が事例を正確に分類する能力が高いものとし，それを含むかどうかで分割する操作を再帰的に行っていく(EBONSAI は BONSAI と同様にシークエンシャルパターン，正規パターンを扱うことができるが，本稿では，分析の過程で部分文字列のみを利用しており，説明の簡便化のため，部分文字列 α として記述している).

EBONSAI は BONSAI と同様に alphabet indexing というメカニズムを内包している．これは正負事例を特徴付ける所与の文字集合をより小さいサイズの文字列に置き換えることで，探索空間を削減すると同時に，より少ない文字列で解釈可能性の高いルールを抽出することを可能にするものである．具体的には以下のようになる．正負事例集合 P，N からサンプル p，n をランダムに取り出す．アルファベット集合から，ランダムに生成されたより小さなアルファベット集合に元の文字列を変換し，上述した手続きで決定木を生成する．次に P，N を用いて近傍を分類精度が改善されなくなるまで探索し，このプロセスを可能な限り繰り返すことで，より判別能力が高い決定木を出力する．適切な alphabet indexing を利用することによって，分類精度の向上，仮説の単純化を実現できる.

EBONSAI の機能はその出力を見ると理解しやすい．図 3 は，EBONSAI の出力例を示している．対象となる正負事例が G，A，T，C の 4 つの文字で構成される文字列をもっているものとする．EBONSAI は図 3 のマッピングテーブルに基づいて，所与の 4 つの文字列を 0 と 1 の 2 つの文字列に置換する(インデックス化)．変換された正負事例の文字列に対して，決定木の root から抽出された文字列と一致するかどうかが調べられる．例えば “11” という文字列が含まれる場合，yes の矢印に振り分けられ，含まない場合，no に分けられる (*は任意の文字列を指す)．この作業を末端の葉まで繰り返すことで，与えられた事例は正例 (pos) か負例 (neg) に分類することができる．このように EBONSAI は少数の変換された文字列を用いることで，よりシンプルで予測力のある決定木を生成する．EBONSAI は主に購買パターン文字列に対して利用されていたため，100 以上の文字で構成された文字列データに対応することが可能である．また，その他の EBONSAI の改良点には以下のようなものがある.

- ビジネスでは複雑な因果関係を扱うために，多様な属性を同時に扱う必要がある．したがって，EBONSAI は複数の文字列属性を利用することができ，また一般的な決定木アルゴリズムのように，文字列属性だけではなくカテゴリ属性，数値属性も同時に 1 つのモデルで扱えるようにした.

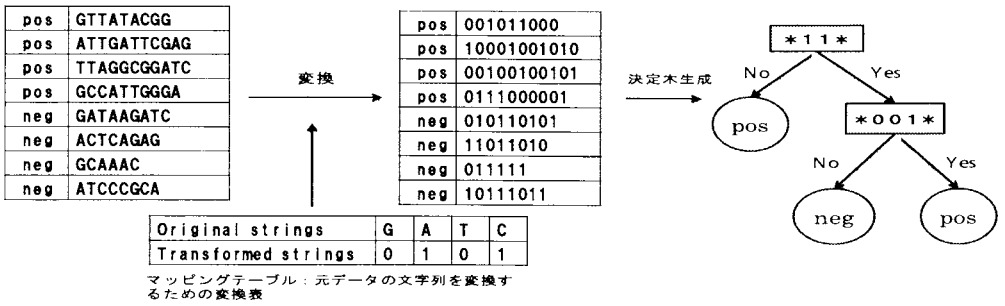

図 3. EBONSAI の出力例.

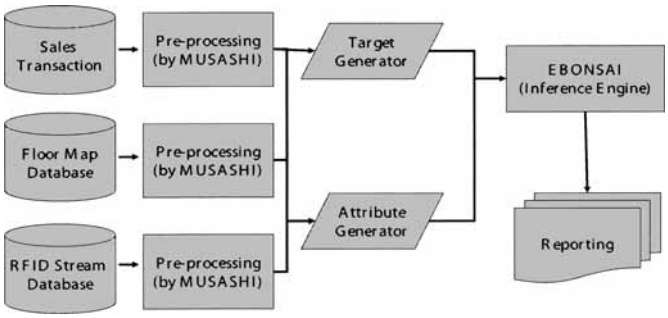

図 4. 顧客動線データからの知識発見システムの概要.

- EBONSAI は XML で記述されたテーブル構造のデータを扱うことができ，オープンソースプラットフォーム MUSASHI を合わせて用いると容易にシステム構築が可能になる.

2.4 システムの概要

図 4 は今回開発した RFID データを用いた知識発見システムの概念図である．元データとして 3 つのデータベースを利用し，それぞれに前処理システムが付随している．前処理システムは，XML データとしてそれらを処理し，次のターゲット属性生成，説明属性の生成のためのシステムに引き渡される．そして，それらのデータを融合し，マイニングエンジンによって，分類モデルが構築される．これらはデータマイニングのオープンソースプラットフォーム MUSASHI (羽室 他, 2005) をベースに構築されている.

これらの主要なサブシステムについて，詳細を説明しよう．本システムでは 3 つのデータベースを利用している．第一は顧客の購買履歴データベースであり，顧客 ID，価格・商品情報などが含まれている．第二に，店舗のレイアウトデータベースである．これには商品データベース，ならび RFID のセンサー位置情報が含まれており，店内の商品の販売位置，ならびに顧客の位置情報を追跡することが可能である．第三に RFID のセンサーログデータベースである．これらをすべて統合すれば，ある顧客がどのように店内を移動し，そしてどの位置にある商品を購入したのかが把握できる.

次にターゲット属性の生成である．ここで開発するのは，様々なデータベースを統合し，最終的に顧客の分類モデルを構築するシステムである．したがって，分類すべき対象となるターゲット属性を上記のデータベースから生成する必要がある．このコンポーネントは購買履歴データベース，RFID のセンサーログデータベースを用い，ユーザーが任意のターゲット属性を生成する．例えば，特定商品の購入者や店舗の優良顧客などが考えられる.

同様に，上述のデータベースから分類モデルで利用する説明属性を生成するコンポーネントが用意されている．ここから特定商品またはカテゴリの購入情報，店内の移動情報などに関する説明属性が生成される．例えば，店内の移動情報から顧客ごとに立ち寄った売場の順序情報がシーケンスの説明属性として生成される.

最後に，マイニングエンジンがこれらのターゲット属性，説明属性から分類モデルを構築する．本システムのマイニングエンジンは EBONSAI をベースに構築されている．したがって，上述したデータベースから生成される数値，カテゴリ，文字列属性を用いた決定木が出力される.

3. 実験結果

3.1 データの説明

本研究で利用したデータは，(株)ボンラパス，(株)博報堂，(株)富士通研究所からご提供いただいたデータをもとに作成した．ここでは，実際の顧客動線データに対して上述のシステムを適用し，ルール抽出の実験を行う．データは，日本の中規模スーパーで収集された顧客動線データを用いた．このプロジェクトでは，顧客が用いるカートに RFID レシーバーを装着し，各売場に RFID タグを設置することで顧客の店内の購買行動を詳細に追跡した．実験は 2006 年 9 月に実施され，顧客動線データのほかに，フロアレイアウト，販売履歴データなども収集された．サンプル数は 216，平均購買時間は 24 分，平均購入数量は約 12 アイテムだった．店内のフロアレイアウトは 7 つのエリアに別れ，それぞれのエリアはサブエリアを持ち，サブエリアの合計は 17 であった．

本実験における分析目的は，この論文で提案したシステムを用いて，より多くのアイテムを購入する顧客がどのような店内移動経路の特徴を持っているのかを明らかにすることである．今回はデータの制約もあり，来店時の購入アイテム数を用いており，それぞれの顧客の購買間隔日数や金額デシル(1 ヶ月あたりの購買金額合計)などは考慮できなかった．カイ 2 乗値でより多くのアイテムを購入した HighVolume 顧客(HV)，それ以外の LowVolume 顧客(LV)に分類した．複数のクラスタリング手法(神嶌, 2003)を適用したがそれらの間に大きな差は見られず，HV の人数は 72 人，LV は 144 人で，HV の 1 回来店あたりの平均購入アイテム数は 19，LV は 7.88 であった．

前述のように，本研究では販売エリアへの訪問における「立ち寄り」に焦点を当てる．本研究では立ち寄りの定義として，「ある販売エリアにおいて，顧客が 1 秒間以上，停止している状態」を採用した．停止している状態とは，顧客の移動速度が秒速 30 cm 以下の状態を指す．ある売場内で立ち寄りの状態が 2 回連続で続き，その間の移動時間が 2 秒未満の場合，1 度の立ち寄りとして処理した．ある販売エリアへの滞在時間は，これら立ち寄り時間の合計となる．これらの定義は，実際の顧客動線データをもとに，店舗担当者やメーカーのマーケティングスタッフと検討し，設定した．顧客の平均立ち寄り販売エリア数は 52.19，標準偏差は 32.5 であった．

今回の実験では，説明属性を生成するコンポーネントから出力された文字列属性と数値属性の 2 種類を利用した．文字列属性としては，顧客の販売エリアの訪問パターン文字列(walking)として，エリアを 1 文字で表現した文字列で構成された売場訪問パターン文字列を用意した．

また，数値属性としては各エリア x $(x = a, b, c, f, m, r, v)$ の滞在合計時間の構成比を利用した．顧客 i がエリア x に t_{ix} 秒滞在したとき，顧客 i のエリア x の滞在時間構成比 r_{ix} は，次のように表せる．

$$r_{ix} = \frac{t_{ix}}{\sum t_{ix}} \tag{3.1}$$

各エリアの滞在時間構成比として，7 属性をモデル構築に利用した．同様の手続きでサブエリアについても訪問パターン文字列(s-walking)と滞在時間構成比 r_{is} $(s = 0, 1, \ldots, 17)$ を算出した．

3.2 訪問パターンとしての文字列表現の有用性

本研究の目的は，訪問パターン分析における文字列表現の適用可能性を示すことにある．ここでは訪問パターン文字列属性が重要な情報を含んでいる可能性を指摘していきたい．最初に，上記で用意した滞在時間構成比 r_{ix} を用いたモデルと訪問パターン文字列(walking)を用いたモデルの予測精度を比較した．滞在時間構成比には顧客の訪問パターン，順序に関する情報は含まれておらず，一方，訪問パターン文字列には滞在時間に関する情報が含まれていない．ただ

表 1. 滞在時間構成比と訪問パターン文字列の予測精度比較.

	用いた属性	Overall Accuracy	Precision(HV)	Recall(HV)	F-Measure
Logit	r_{ix}	0.727	0.623	0.615	0.619
C4.5	r_{ix}	0.75	0.62	0.795	0.697
EBONSAI	$walking$	0.764	0.615	0.923	0.738

し実務家やマーケティング研究者の意見によると，売場への滞在時間や滞在時間構成比は顧客の商品購入に関係があることが推測される．両者を用いたモデルの精度を比較することで，訪問パターン文字列に何らかの重要な情報が含まれているかどうかを推測できると考えた．滞在時間構成比のモデルとしては，ロジスティック回帰，C4.5，訪問パターン文字列には EBONSAI を用い，交差検証(10 fold)の平均として Overall accuracy，Precision，Recall，そして F-Measure (Witten and Frank, 2000)を算出した．EBONSAI は C4.5 と同じ枝刈のアルゴリズム(Quinlan, 1993)を採用しており，両者ともパラメーターは cf=0.25，葉あたりの最小事例数 15 で実験を行った．実験には WEKA (Witten and Frank, 2000)を利用した．表 1 からわかるように，訪問パターン文字列を用いた EBONSAI のモデルは滞在時間構成比の 2 つのモデルとほぼ同等もしくは高い分類精度を持っていた．データの制約上，あらゆる属性を生成し比較することはできないが，訪問パターンには滞在時間構成比のように，顧客の購入量に関する重要な情報が含まれていると推測できる．

現実に重要な知識を発見しようとする場合，手法の予測精度を評価するだけでなく，専門家にとって，抽出されるルールが新しい知見をもたらすのか(元田 他, 2006)，そして実際にビジネスアクションを生み出すことができるのか(矢田, 2004b)という観点から評価することが重要である．訪問パターン文字列の有用性を示すためにも，専門家にどのような知見をもたらすのかを明らかにすることは重要である．そこで滞在時間構成比，訪問パターン文字列の属性を用いて構築された決定木を用い，滞在時間構成比から得られたルールと訪問パターン文字列を追加して得られたルールを専門家に比較してもらい，抽出される知見を検討した．専門家の解釈可能性を高めるために，枝刈のパラメーターは cf=0.10，葉あたりの最小事例数 30 で算出した．以下で専門家らによるルールの解釈を要約する．

図 5 の滞在時間構成比を用いた決定木は，専門家の従来の常識と一致するルールになっていた．トップノードは魚売場の滞在時間構成比が 0.23 以上であると HV 顧客に分類されるというルールである．従来からスーパー業界では生鮮三品(青果，鮮魚，精肉)の重要性は広く知られ，特に価格によらない集客商品として鮮魚は店舗にとって重要な意味を持つものと言われてきた．彼らによると鮮魚売場はその店舗の顧客忠誠度を惹きつけるもっとも重要な売場だと考えられていた．こうした常識が HV 顧客の滞在時間構成比の高さとなって現れていると解釈できる．次のノードでは，賞味期限の短い日配品が含まれる定番棚(b)が 0.1 未満の場合，HV 顧客に分類されている．日配品の構成比が低いということは，必要なものだけではなく，多様な商品を品定めしているのではないかと思われる．そして最終ノードでは，肉売場(m)が 0.17 以上であると HV 顧客に分類されている．これも上述の鮮魚と同様に，顧客が重視するカテゴリであることから，常識と一致すると解釈された．

次に滞在時間構成比と訪問パターン文字列を用いて，EBONSAI による分類モデルの構築を行い，抽出されたルール(図 6)を専門家に評価してもらった．トップノードは図 5 と同じであるが，次のノードには訪問パターン文字列が使われ，V (野菜売場)から F (鮮魚売場)に向かう訪問パターンがない場合，LV 顧客に分類されている．そして最後のノードは，肉売場(m)が 0.035 以上であると HV 顧客に分類されていた．得られたルールで特徴的なのは，野菜売場か

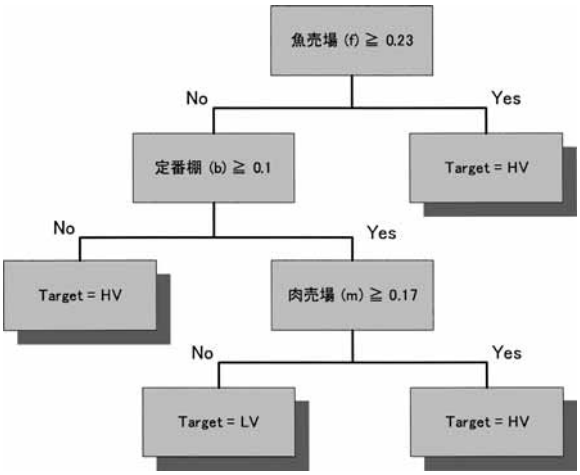

図 5. 滞在時間構成比を用いた決定木の一例.

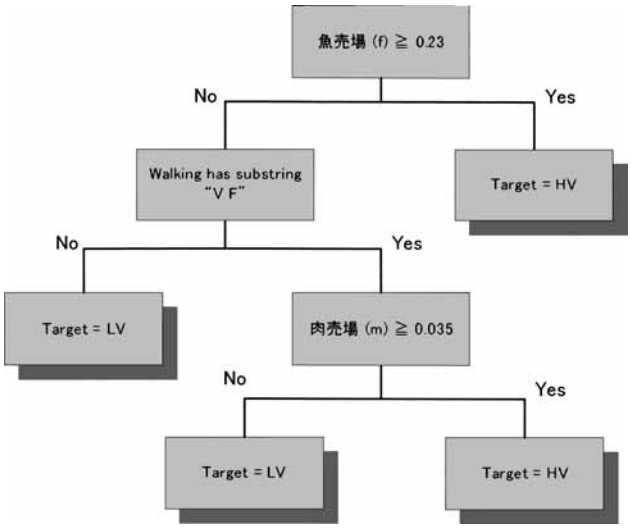

図 6. 滞在時間構成比と訪問パターン文字列を用いて EBONSAI によって抽出されたルール.

らどの売場(定番棚もしくは魚売場)へ移動するのかが購入アイテム数，つまり HV 顧客と LV 顧客の分岐点になっていることを示している点である(図 7 を参照)．従来，店舗の担当者らは魚売場での滞在時間構成比が大きな顧客ほど買上げ点数が高くなるという単純な仮説を持っていた．しかしながら，この実験では，販売エリア訪問パターンにも重要な特徴が見られた．そして鮮魚売場に誘導できたとしても，隣接する肉売場で立ち寄らない場合，つまり鮮魚売場から定番棚エリア(c)へ入ってしまう，または肉売場を通過してしまう場合，LV 顧客になるということは，以前の野菜売場や鮮魚売場において，十分な需要喚起が達成できていないのではないかと推測された.

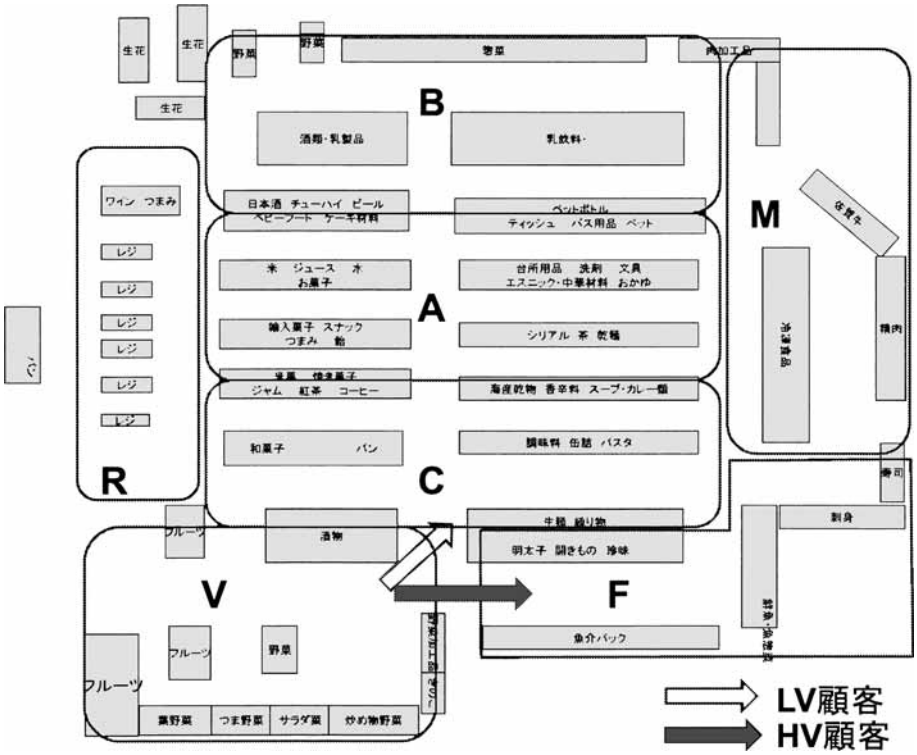

図 7. HV 顧客と LV 顧客の売場間移動.

店舗担当者などの実務家が最も関心を持ったのは，野菜売場から鮮魚売場への移動に関するルールであった．店舗の担当者とこれらの知見を議論したところ，HV 顧客の多くが野菜売場で何らかの刺激を受け，購入商品(鮮魚)を決定しているのではないかという仮説を得ることができた．また彼らは野菜売場での鮮魚の需要喚起が，顧客の店内での購買意欲を規定するのではないかと考えた．これらの仮説に基づくと，小売店にとって，野菜売場から魚売場への動線に顧客をどうやって誘導するかが，重要な店内レイアウト上の課題になると考えられる．

最後に訪問パターン文字列という知識表現の有用性として専門家から指摘されたものは，以下の 2 点にまとめられる．

- 店内レイアウトをクロスセクションの観点から検討する基礎データを提供しうる．店舗では基本的に各セクションの商品担当者(主にバイヤー)が売上実績をもとに売場構成や品ぞろえを決定する．訪問パターンのルールは，セクションの垣根を超え，複眼的な視点から顧客の店内行動をもとに売場作りを行う基礎データとなる．ケースでは，野菜売場から鮮魚売場へ誘導する「売場間のストーリー」が必要であることが検討された．
- 抽出された訪問パターンはビジネスアクションの実施場所への示唆を生むことがある．例えばケース中のルールであれば，野菜売場から鮮魚売場への移動を促進するために，そのエリア間で関連メニュー提案などの実施が考えられる．また，野菜売場と定番棚(c)の境界では，顧客移動を抑制するため，野菜売場関連の販売促進策ではなく，定番棚(c)のセール商品を設置することが検討された．

このようなビジネスへの示唆を提供できることから，訪問パターン文字列は顧客の店内行動の情報を含んでおり，顧客動線分析における重要な知識表現であると考えられる．

3.3 文字列解析手法 EBONSAI の技術的課題

上述の検討から，訪問パターン文字列が顧客の店内行動に関するリッチな情報を含んでおり，したがってそれを直接扱うことができる文字列解析手法は有用な知見を創出する可能性があると考えられる．しかしながら，本稿で文字列解析手法として利用した EBONSAI には，顧客動線分析に適用するために，解決すべき問題が残されていることがわかった．ここでは EBONSAI，文字列表現の問題に焦点をあて，今後の課題を明らかにする．

3.3.1 顧客動線分析における alphabet indexing

最初に EBONSAI の alphabet indexing について検討を加えたい．EBONSAI において indexing は，単に探索空間の削減だけではなく，ルールの意味解釈を容易にするという効果を持つ．例えば，ブランドスウィッチパターンへ EBONSAI を適用した場合，同じメーカーから発売されているブランドや同じ味，類似のターゲットを持つブランド（低価格帯の商品など）が indexing によって，1 つの文字に置き換えられることが多い．これによって，出力されるルールが単純化され，ルールの解釈可能性が高くなるという効果を持つ．しかし，顧客動線分析においては，ルールの解釈可能性という点で indexing が十分に機能しない．なぜなら複数の売場が 1 つの文字に置き換えられると，その置き換えられた売場グループに特定の意味を見つけ出すことが困難なためである．実際に専門家との議論において，indexing はむしろ混乱を招き，上述の図 6 のように，EBONSAI を用いたルール抽出では indexing を利用しない決定木だけが専門家から関心を集めた．顧客動線分析では，目的変数に大きな影響を持つ特定の経路を見つけることに多くの専門家が興味を持っているため，indexing を利用しない，もしくはインデックスのサイズをなるべく大きくとる方法を検討すべきである．

しかし，インデックスのサイズを大きくすることは，計算時間の増大を招く．図 8 は EBONSAI のパラメーターであるインデックスのサイズとルールを出力するまでの計算時間の関係を示したものである．この実験では，説明属性に店内が 17 の販売エリアに分割されているサブエリアの訪問パターン文字列（s-walking）を採用した．EBONSAI のインデックスのサイズはパラメーターで設定する仕様になっており，デフォルトは 2 である．予測精度という観点からみれば，多くの場合，インデックスサイズが 2 もしくは 3 のときに高い精度が得られることが多い．またインデックスサイズが小さいほど計算時間は短縮される傾向を持つ．図 8 のグラフにおいて，インデックスのサイズ 7 から計算時間が長くなるが，それ以降の極端な伸びはなかった．その理由としては，極端に訪問頻度の低いサブエリアが多く含まれていたためと考えられる．もし巨大なサンプル数が収集可能で，すべてのサブエリアが一定以上の訪問頻度をもつような店舗の場合，インデックスサイズの増大が計算時間の極端な増大を招く可能性もある．また，現在の EBONSAI の最大のインデックス数は 9 であるため，今後，さらに大きなサイズのインデックスが扱えるよう改良する必要がある．そして indexing 以外の新しいアプローチによる探索空間の削減方法について，検討する必要があるだろう．

3.3.2 顧客動線分析におけるシーケンシャルルール

EBONSAI は部分文字列だけではなく，正規表現などで表されるようなシーケンシャルパターンを含む決定木を生成する（Hamuro et al., 1998; Yada et al., 2007）ことが可能である．しかしながら，専門家はシーケンシャルパターンを含んだルールに関心を払わず，連続した移動販売エリアのルール，つまり部分文字列で構成された決定木を生成するように要求した．シーケンシャルパターンの場合，売場訪問パターン文字列は売場の相対的な訪問順序を示しており，連

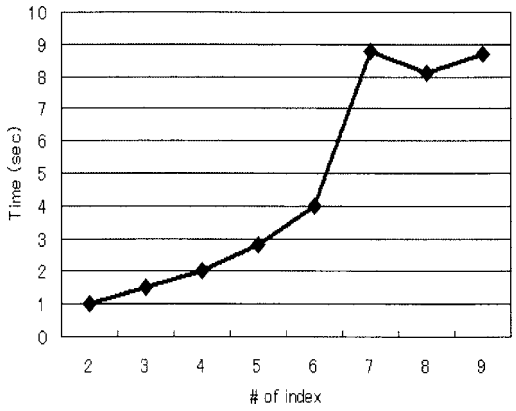

図 8. インデックスサイズと計算時間.

続した訪問とは限らない．専門家は連続した売場訪問に関する情報に有用性を認めているが，シーケンシャルパターンは解釈可能性，ビジネスアクションへの示唆という点で問題があるという評価であった．

ケースで抽出されたルールが比較的短い部分文字列であることを考慮すると，EBONSAI が訪問パターン文字列を解析するための唯一の方法ではない．例えば，LCMseq (Uno et al., 2004) のような頻出パターンの列挙アルゴリズムを用いて抽出されたものを既存の C4.5 などの決定木で解析する方法も考えられる．今後，EBONSAI だけではなく，他の文字列解析手法を比較検討し，データや環境，分析ニーズに合った解析手法はどのようなものかを明らかにする必要があると思われる．

3.3.3 滞在時間情報の欠落

顧客動線分析において，本稿で扱った文字列という知識表現は大きな課題が残る．最も重要な問題は，訪問パターンを文字列に変換する際に，顧客の店内行動に関する時系列情報が大幅に消失していることである．例えば，売場間の移動時間やある売場への滞在時間といった重要な情報が文字列の知識表現の中に反映されていない．こうした問題の解決には，グラフ構造データ (Yada et al., 2006) の導入といった新たな知識表現の導入が考えられる．グラフ構造データであれば，販売エリアの訪問パターンだけでなく，売場間移動時間や滞在時間などの時系列情報を含めることが可能である．今後，こうした他の知識表現に対して顧客動線分析の有用性の観点から検討を加える必要がある．

4. むすび

本研究では RFID を用いて得られた顧客動線に関するストリームデータに対して，既存の文字列解析手法を適用し，顧客の購買行動に関する知識発見を試みた．我々は顧客動線データの中で顧客の売場への立ち寄りに注目し，販売エリアの訪問パターンを文字列で表現することによって，膨大なストリームデータを効率よく扱うことを提案した．より多くのアイテムを購入する HV 顧客には野菜売場から魚売場へと移動する特徴的な訪問パターンが見られた．こうして得られた仮説は新規性が高く，示唆に富むものであった．またこの実験を通して，顧客動線分析における既存の文字列解析手法の課題を理解することができた．

しかしながら，本研究に残された課題は多い．本論文では顧客動線分析への文字列解析技術の応用可能性を示すことができたが，その一般化を十分に証明できたとは言えない．例えば本論文では，公開されていない1つの業態(中規模スーパーマーケット)での実験データのみを用いている．したがって，異なる小売業態(薬局，ホームセンター，ショッピングセンターなど)への適用可能性など，文字列解析技術の顧客動線分析への一般的な有用性を明らかにすることはできなかった．またデータの制約上，過去の顧客の購買行動など顧客特性を十分に考慮した分析ができなかった．今後，多様な業態，顧客グループ間比較などへ実験を拡張し，適用可能性の一般化を検討する必要がある．

参 考 文 献

Arikawa, S., Miyano, S., Shinohara, A., Kuhara, S., Mukouchi, Y. and Shinohara, T. (1993). A machine discovery from amino acid sequences by decision trees over regular patterns, *New Generation Computing*, **11**, 361–375.

有村博紀(2005). 大規模データストリームのためのマイニング技術の動向，電子情報通信学会論文誌 D-1, **J88-D-1**, 563–575.

Guadagni, P. M. and Little, J. D. C. (1983). A logit model of brand choice, calibrated on scanner data, *Marketing Science*, **2**, 203–238.

Gupta, S. (1988). Impact of sales promotions on when, what, and how much to buy, *Journal of Marketing Research*, **25**, 342–355.

Hamuro, Y., Katoh, N., Matsuda, Y. and Yada, K. (1998). Mining pharmacy data helps to make profits, *Data Mining and Knowledge Discovery*, **2**, 391–398.

Hamuro, Y., Kawata, H., Katoh, N. and Yada, K. (2002). A machine learning algorithm for analyzing string patterns helps to discover simple and interpretable business rules from purchase history, *Progress in Discovery Science*, **LNAI 2281**, 565–575.

羽室行信，加藤直樹，矢田勝俊，鷲尾　隆(2005). 大規模ビジネスデータからの知識発見システム：MUSASHI，人工知能学会誌，**20**, 59–66.

Hirao, M., Hoshino, H., Shinohara, A., Takeda, M. and Arikawa, S. (2003). A practical algorithm to find the best subsequences patterns, *Theoretical Computer Science*, **292**, 465–479.

神嶌敏弘(2003). データマイニング分野のクラスタリング手法(1)，人工知能学会誌，**18**, 59–65.

経済産業省(2005).『日本版フューチャーストア・プロジェクト』について，News Release, 1–10. (http://www.meti.go.jp/press/20051108001/20051108001.html)

Larson, J. S., Bradlow, E. T. and Fader, P. S. (2005). An exploratory look at supermarket shopping paths, *International Journal of Research in Marketing*, **22**, 395–414.

元田　浩，津本周作，山口高平，沼尾正行(2006).『データマイニングの基礎』，オーム社，東京.

Quinlan, R. (1993). *C4.5:Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, California.

Sorensen, H. (2003). The science of shopping, *Marketing Research*, **15**, 30–35.

Shimozono, S., Shinohara, A., Shinohara, T., Miyano, S., Kuhara, S. and Arikawa, S. (1994). Knowledge acquisition from amino acid sequences by machine learning system BONSAI, *Transaction of Information Processing Society of Japan*, **35**, 2009–2018.

Uno, T., Kiyomi, M. and Arimura, H. (2004). LCM ver.2: Efficient mining algorithms for frequent/closed/maximal itemsets, *Proceedings of IEEE ICDM'04 Workshop FIMI'04*, 1–11.

Witten, I. H. and Frank, E. (2000). *Data Mining*: *Practical Machine Learning Tools and Techniques with JAVA Implementation*, Morgan Kaufmann, San Francisco, California.

矢田勝俊(2004a). マーケティングにおけるデータマイニングの利用，人工知能学会誌，**19**, 376–377.

矢田勝俊(2004b). 『データマイニングと組織能力』, 多賀出版, 東京.

Yada, K., Washio, T. and Motoda, H. (2006). Consumer behavior analysis by graph mining technique, *New Mathematics and Natural Computation*, **2**, 59–68.

Yada, K., Ip, E. and Katoh, N. (2007). Is this brand ephemeral? A multivariate tree-based decision analysis of new product sustainability, *Decision Support Systems*, **44**, 223–234.

Proceedings of the Institute of Statistical Mathematics Vol. 56, No. 2, 199–213 (2008) 213

Path Analysis in a Supermarket and String Analysis Technique

Katsutoshi Yada

Faculty of Commerce, Kansai University

This paper presents the availability and usefulness of a string analysis technique for developing useful rules to determine customers' visiting patterns in sales area. It focuses on stationary states of customers in certain sales areas in a store. We apply a string analysis technique, EBONSAI, to sales area visiting patterns to effectively deal with a huge stream of data. Experiments were conducted to extract useful rules and findings about characteristics of sales area visiting patterns and we discuss problems remaining in existing string analysis techniques.

Key words: Supermarket, marketing, RFID, string analysis technique, EBONSAI.