ベイジアンネットワークを用いた消費者行動モデルの構築実験

左　毅，矢田　勝俊

本稿では，RFID 技術を用いた店舗内行動の把握と購買行動予測への応用について，最新の研究を報告する．既存研究の多くは店舗内全体での移動行為や購買行為に注意を払っていたのに対して，本研究は顧客の異質性と売場の商品特性の違いを前提にした売場滞在時間と購買行動の関係を明らかにするモデルを構築している．われわれは確率的グラフィカルモデルであるベイジアンネットワークを用いて，滞在時間中に行われる購買行動に関する意思決定プロセスの定量的な分析を行った．次に顧客の購買傾向の異質性を考慮するために，購買履歴から特徴量としての独立変数を作成し，ベイジアンネットワークを構築した．ベイジアンネットワークは離散変数しか扱えないため，クラスタリング・アルゴリズムを用いて連続変数を離散化する必要がある．実験において提案手法の性能を評価するため，滞在時間と購買傾向に関する説明変数の最適なクラスター数を検討し，ほかの典型的な予測モデルと比較してベイジアンネットワークの精度が最も高いことを示した．提案モデルは顧客の購買傾向の違いを考慮し，ROC 曲線を用いてモデルの分類精度を最大化する最適な購入意思決定の閾値を特定した．

キーワード：購買行動予測，店舗内行動，ベイジアンネットワーク，クラスタリング，ROC 曲線

1. はじめに

近年，RFID 技術はスーパーマーケットで活用され始めており，店舗内の顧客の行動を追跡し，データとして蓄積することを可能にしている．こうした店舗内行動に関する詳細なデータはコンピュータサイエンスだけではなく，ビジネス分野など多くの研究者の興味を集めている．従来，顧客の店舗内行動を把握するためには，観察者がそれぞれの顧客を物理的に追跡し，記録する必要があった．RFID 技術はこれらを代替し，自動的に大量の顧客の店舗内行動をデータ化し，蓄積することで，多くのビジネス上の知見を提供している．本稿の目的は，店舗内行動に関する知見に重点を置き，顧客の購買行動に対する理解を深めることである．われわれが過去に行った研究 [1] ではベイジアンネットワーク (BN) を採用し，滞在時間による購買意思決定の変化を分析している．当該研究領域における本稿の貢献として三つの点を示すことができよう．第一に，店舗レイアウトの離散化（最小化）に専門知識を導入し，商品カテゴリ（パン売場）レベルのモデル化を行うことで，単一ブランドやカテゴリ内の商品群に対する顧客の行動を詳細に理解することが可能になったことである．この改善によって，商品配置や品揃えサービスの向上など，より実践的な含意を導き出すことができる．第二に，モデルには年齢や性別などのデモグラフィック要因ではなく，購買履歴から得られる購買傾向を利用した点にある．購買履歴は商品に対する顧客の態度とロイヤリティを示す計測可能かつ累積的な情報である．これにより消費者の態度と行動の両方の視点から購買行動の意思決定プロセスを検証することができる．第三に，実務に必要な前処理の実例として，独立変数の離散化の手続きを示している点である．ベイジアンネットワークを利用する場合に限らず，実務では前処理で独立変数を離散化しなければならないことが多い．本稿では，K 平均法を用いて学習データを離散値に変換し，BIC [2] を基準に最適なクラスター数を推定している．われわれは提案モデルの性能を検証するため，複数の典型的な予測モデルと比較している．購買行動予測の精度を高めるためには，少数グループに購入データが偏在している場合でも，精密な予測ができることが望ましい．本稿では，それぞれ陽性データと陰性データの的中率を示す感性 (sensitivity) と特異性 (specificity) という複数の基準を用いて予測モデルの性能を評価した．感性の増加関数と特異性の減少関数である ROC 分析 [3] を用いることで，意思決定の閾値を調整するために ROC 曲線を生成した．

本稿の構成は以下のとおりである．2 節では，店舗

さ　き
名古屋大学未来社会創造機構
〒 464–8601 愛知県名古屋市千種区不老町
zuo@coi.nagoya-u.ac.jp, zuo@nagoya-u.jp
やだ　かつとし
関西大学商学部
〒 564–8680 大阪府吹田市山手町 3–3–35
yada@kansai-u.ac.jp

内行動に関する過去の研究成果や近年の研究テーマを紹介する．3 節では，筆者らの過去の研究を中心に当該領域の本質的な問題について説明する．4 節では，提案モデルで用いられるベイジアンネットワークと，二つの方法論的課題を提示する．5 節では，実験内容や性能評価の結果を示す．最後に結論や今後の研究の方向性を示す．

2. 関連研究

2.1 店舗内行動に関する初期の経験的アプローチ

店舗内行動に関する初期の研究 [4, 5] においては，観察者または調査員が必ず必要となる．店舗レイアウトにおける顧客の行動を追跡し，顧客の購買行動および店舗内行動を詳細に記録するのである．コストが高く，作業が膨大であることに加えて，そういった実験を現代のスーパーマーケットにおいて実行することは，ほとんど不可能である．観察や計測によって，顧客の通常のショッピングの妨げとなってしまい，スーパーマーケットのマネージャーが実験に拒否反応を示すからである．

消費者行動研究の分野において，アンケートやインタビューなどの方法論が開発され，思考や感情がどのように購買行動に影響を与えるのかを抽出することが可能となっている．しかしながら，依然として顧客の店舗内行動の直接的な観察には光が当てられていない．しかも，質問および回答の選択肢は研究者の手によるものであるので，研究者の主観的なレンズを通して顧客が回答しているという可能性もありうる．

一方，顧客が回答した結果というのは，自己申告に基づいた調査結果である．顧客の表現能力，そして記憶能力には限界があるため，回答データは不完全で，トレーニングデータとして適用するのは困難であることも少なくない．顧客の意思決定プロセスを正確に理解し，効果的なインストアマーケティングを策定するためには，顧客の店舗内行動を定量的に測定し，さまざまなデータと統合して分析する必要があると言える．

2.2 顧客動線研究

2000 年代以降，顧客の店舗内行動の新しい観察方法は，小売り産業の分野における技術革新であった．スーパーマーケットにおける顧客の移動データを記録するために RFID 技術が導入され，顧客動線データは，小売業者およびメーカーを将来の成功に導くことのできる重要な戦略的資源として，ますます注目を集めている．

店舗内行動の最初の顧客動線研究（PathTracker® システムによって収集されたデータ）は，Sorensen [6] によって，アメリカ西部のスーパーマーケットで実施された．ショッピングカートおよび買い物かごの底に RFID タグが取り付けられて，それぞれの RFID タグが店舗内の位置を補足するためのシグナルを発し，レセプターが受信する．ショッピングの最後には，それぞれの移動経路と売買取引とのマッチングが行われ，顧客動線データとしてデータベース内に記録される．ソーレンセンの顧客動線データに基づき，Larson et al. [7] は食料品店内における移動経路の特徴を抽出し，14 種類の標準的な移動経路に集約した．この研究によって顧客動線データの存在が多くの研究者の注目を集めるきっかけとなり，その後の当該領域の研究に大きな影響を与えることになる．

もう一つの代表的な研究は，Hui et al. [8] によって提唱されたものである．彼らは同じ PathTracker® システムをアメリカ東部のスーパーマーケットに設置し，顧客動線データを分析するために，ベイズの統計的推論を使用して，統合的な確率モデルを提案した．クラスタリング・アルゴリズムを使用した移動経路の分析である Larson et al. の研究に対して，ヒュイらは顧客動線データに関する消費者行動（訪問，ショッピング，購買）についての，3 種類の仮説の検証を行った．彼らの論文では，三つの状況要素（時間要因，ライセンシング要因，社会要因）が，消費者行動に対して与える影響について論じられている．しかしながら，依然として顧客動線データは購買行動の予測に適用されることはなかった．購買行動の予測ができれば，顧客の店舗内行動から，小売業者やそのマネージャーに対して，意義深く価値のある知見を提供することができるはずである．

矢田の研究 [9] においては，RFID 技術によるデータ収集の正確性が改善され，位置シグナルの発生頻度が 5 秒間隔から 1 秒間隔へと引き上げられた．この改善によって，店舗内行動の観察調査に対する詳細な説明（たとえば，棚および商品の特性抽出）が可能となった．この研究においては，多くの商品を購入する顧客の訪問パターンを探るために，文字列解析のテクニックを適用して，買物の際の移動経路の分析が行われた．そして，ターゲット顧客の訪問パターンのルールを抽出し，購買行動の説明とともに店舗にとって重要な知見を明らかにした．

このように顧客動線データから，二つの主要な情報（訪問パターンおよび買物時間）を抽出することができるのであるが，上記の研究では訪問パターンの有用性が示されていた．買物時間に焦点を合わせた場合，矢

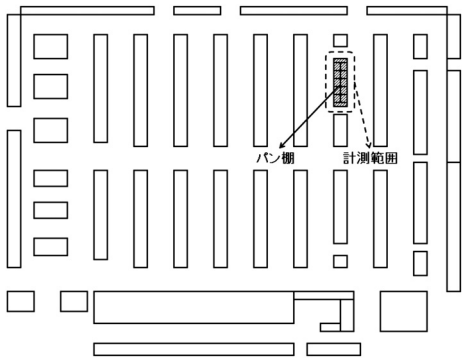

図 1　パンカテゴリーにおける商品の計測エリア

田の研究においては，顧客を「大口」客または「小口」客に分けるための分類属性として，それぞれの売場における買物時間の比率が使用されている [9]．しかしながら，「顧客が店内で多くの時間を過ごすほど，特定のゾーンにいるときにショッピングモードに入る可能性が高くなる」ということが Hui et al. の研究において証明されており [8]，買物時間が購買意図においてどのように変化を引き起こすのかについてのさらなる解析が必要である．とりわけ，特定のエリアにおいて過ごした時間を訪問パターンに組み入れるべきであるということも，Yada によって提唱されている [9]．

3.　前処理

3.1　実験ターゲットの選定

以前の研究においては [1]，鮮魚が日本国内のスーパーマーケットにとって最も重要なカテゴリであるため，鮮魚の販売エリアが実験のターゲットとして選定された．本実験では，パンのカテゴリを実験ターゲットとして選択した．現代の日本では伝統食であるご飯よりも，パンが朝食において主役となっており，重要な商品カテゴリとされる．

3.2　計測エリアの改善

過去の研究で示されたように，滞在時間は，購買の意思決定がなされるまでに費やされた時間だけでなく，購買後の時間も含まれるものである．滞在時間と購買の意思決定の関係をより明確にするために，計測エリアを広いカテゴリのセクションではなく，小さな範囲（ブロック）へと狭めた．この改善によって，一つの商品ブランドまたは同一のカテゴリに属している一連の商品における，顧客行動の直接的な観察を行うことが可能になる．

上記の 3.1 節において説明したように，パンのカテゴリが実験ターゲットとして選定され，図 1 において示しているような島形陳列に配置されている．われわれは，棚からの距離が 1 メートル以内であるこの島の周囲を計測エリアとして設定した．RFID タグによって発せられる位置シグナルがこの範囲の内側であった場合にのみ，滞在時間として計測することができるものとする．この改善の目的は，あてもなく歩き回っている顧客の影響を減少させて，計測可能な滞在時間を，購買意思決定のために費やされた実際の時間に近づけることである．

4.　方法論

4.1　ベイジアンネットワーク

ベイジアンネットワーク (BN) とは，有向非巡回グラフを通じて，一連の確率変数およびその条件付き依存関係を表す確率的グラフィカルモデルである．ベイジアンネットワークの確率論はベイズの定理に基づいていて，観察された二つのイベントが A → B のような関係性を有しているのであれば，以下のように表すことが可能である．

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}. \tag{1}$$

$\Pr(A)$ および $\Pr(A|B)$ は，それぞれイベント A の事前確率，およびイベント B が発生した下で，イベント A が発生する事後確率を示している．$\Pr(B|A)$ は，尤度関数を示している．分母 $\Pr(B)$ は，イベント A のすべての状態における周辺分布を示している $\Sigma_i \Pr(B|A = a_i)\Pr(A = a_i)$ と等しい．式 (1) を使用することによって，$\Pr(A|B)$ は事前確率 $\Pr(A)$ と尤度 $\Pr(B|A)$ のかけ算で求められる．一般的には最尤法によって，一連の変数に対する，尤度関数を最大化するモデルパラメータの一連の値が選定される．

4.2　変数の離散化

ベイジアンネットワークを利用する場合，離散変数にしか適用することができないので，連続変数である購買傾向および滞在時間を離散化する必要がある．

4.2.1　購買傾向

購買傾向の離散化のために，過去の購買履歴に基づいて顧客の分類が行われた．われわれは，購買傾向のクラスターを次のように定義した．

$$\{B_1, B_2, \ldots, B_L\} \tag{2}$$

ここで，B_l および L は，それぞれ一つの顧客クラスター，およびクラスターの総数を示している．購買傾向は，K 平均アルゴリズムによって分割される．

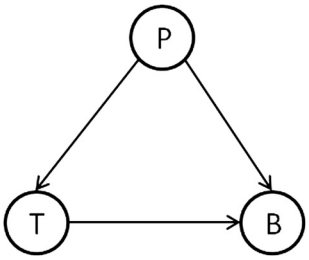

図 2　購買行動予測のベイジアンネットワーク

4.2.2　滞在時間

滞在時間の離散化のために，K 平均アルゴリズムも適用して，M 個のクラスターにおける滞在時間を次のように離散化させた．

$$\{T_1, T_2, \ldots, T_M\} \tag{3}$$

ここで，T_m は滞在時間のクラスターを表す．

4.3　ベイジアンネットワークのグラフィカル構造

本研究においては，購買行動 (P)，滞在時間 (T)，購買傾向 (B) という，三つの変数が存在する．われわれの過去の研究とは対照的に，年齢に代わって購買傾向という新しい態度に関する変数を導入している．デモグラフィック属性とは異なり，購買傾向は自己申告に基づく情報ではなく，商品に対する顧客の認知程度を示すための，計測可能で累積的な要因である．滞在時間との相乗効果によって，認知および行動に関する両方の視点から，購買行動の意思決定プロセスを論証することが可能である．

顧客の購買傾向は，$B \to T$ および $B \to P$ のように，滞在時間および購買行動に対して，直接的に影響を及ぼすことが考えられる．さらに，滞在時間は，$T \to P$ のように，購買行動に対して直接的に影響を及ぼすことが考えられる．購買行動は目的変数であると考えられているので，ベイジアンネットワークを図 2 のように構成すべきである．

式 (1) によると，顧客が購買に至る状態の確率は，次のように推定することが可能である．

$$\begin{aligned}&\Pr(P=1|T,B)\\ &=\frac{\Pr(P=1)\Pr(T|P=1)\Pr(B|T,P=1)}{\sum_{P\in\{0,1\}}\Pr(P)\Pr(T|P)\Pr(B|T,P)}\end{aligned} \tag{4}$$

分子は，購買状態 ($P=1$) および非購買状態 ($P=0$) における周辺分布である．

4.4　クラスター数の選定基準

ベイジアンネットワークは離散変数しか適用することができないので，その問題はクラスタリング・アルゴリズムを使用することによって解決することが可能である．ただし，適切なクラスター数をどのようにして決定すればよいかという新しい問題を考慮に入れる必要がある．とりわけ，二つよりも多くの変数が存在する場合には，一連の変数から最適なクラスター数の組み合わせをどのようにして見つけるのかという，最適化の問題が生じる．

本稿ではクラスター数の選定のためには BIC（ベイジアン情報基準）が採用され，BIC の最小値が最適なクラスター数を決定するための基準となっている [2]．BIC の評価関数の定義は，次のように，第一項としての誤差項，および第二項としてのペナルティ項から構成されている．

$$BIC = n \cdot \ln(\sigma^2) + (L+M) \cdot \ln(n). \tag{5}$$

σ^2 は，残差を示している．購買行動は二値変数であるので，われわれは（1 − 的中率）を誤差項として使用している．L および M はそれぞれ購買傾向および滞在時間のクラスター数，n はサンプルデータのサイズを示している．式 (5) は，的中率の減少関数およびクラスター数の増加関数である．クラスター数をペナルティ項とすることによって，式 (5) の低いほうの値が選択される．

4.5　ROC 曲線

2 クラスの分類問題において，陽性データが少数の場合，トレーニングデータにおける的中率を最大化するだけでは不十分である．本稿においては，モデルの性能は複数の基準に基づいて評価が行われる．その基準が感性および特異性であり，感性は陽性データに関する的中率，特異性は陰性データに関する的中率を示している．

ROC 曲線分析においては，感性の増加関数および特異性の減少関数を使用し [3]，意思決定の閾値の調整のために，ROC 曲線を生成させている．モデルの感性および特異性として，0 から 1 までの間で真陽性率対偽陽性率が与えられる．AUC（曲線下の面積）が 1 に近づけば近づくほど，モデルの性能が高くなることを示している．

感性および特異性は，次のように表される．

$$\text{感性} = \frac{TP}{TP+FN}, \tag{6}$$

$$\text{特異性} = \frac{TN}{FP + TN} \tag{7}$$

TP, FN, TN, FP は，それぞれ真陽性のデータ数，偽陰性のデータ数，真陰性のデータ数，偽陽性のデータ数を示している．

5. 実験

5.1 変数の初期設定

実験は 2009 年 5 月 11 日から 2009 年 6 月 15 日までの間に実施され，パンのカテゴリにおいて合計で 1,155 件の移動経路が抽出された．モデルの検証方法としてはホールドアウト法を使用するので，収集したデータを二つのグループに分けた．トレーニングデータとして 2009 年 5 月 11 日から 2009 年 6 月 10 日まで（924 のサンプルデータが含まれている）のデータを，テストデータとして，2009 年 6 月 11 日から 2009 年 6 月 15 日までのデータ（231 のサンプルデータが含まれている）を使用した．

5.1.1 購買行動

目的変数は，二値変数 0/1 として定義される購買行動である．0 は対象カテゴリでの非購買状態，1 は購買状態をそれぞれ示している．購買行動を検証するために，それぞれのショッピングの移動経路と POS データを用い，購買行動の事前確率 $\Pr(P = 1)$ は，31.95% と計算される．

5.1.2 滞在時間

滞在時間は，行動要因として使用される説明変数の一つである．図 1 において示されている計測範囲に顧客が入った場合にのみ，位置情報がデータ収集のターゲットとして認められる．それぞれのターゲットの合計経過時間が，滞在時間として算出される．

5.1.3 購買傾向

もう一つの説明変数は，認知程度に関する要因として使用される購買傾向である．購買行動および滞在時間とは対照的に，この変数には RFID 実験前の直近 3 カ月間という独立したデータ収集期間が用意されている．2009 年 2 月 11 日から 2009 年 5 月 10 日までの間に，過去の POS データから購買傾向が抽出される．また，購買傾向は商品に対する顧客の認知程度を示すための累積的な要因でもある．滞在時間との相乗効果によって，認知および行動に関する両方の視点から，購買行動の意思決定プロセスを検証することが可能である．

5.2 クラスタリングの最適化

この節においては，予測の精度に対するクラスター数による影響について論じる．それぞれの説明変数に対して，2 から 10 のクラスター数を設定し，式 (5) を使用してすべての組み合わせにおいて BIC を算出する．図 3 において示されているように，BIC の結果は密度推定によって視覚化できる．X 軸および Y 軸は，それぞれ滞在時間のクラスター数 (M)，および購買傾向のクラスター数 (L) を示している．この図によって，BIC が最小値となるのは，$L = 7, M = 6$ のときであることがわかる．

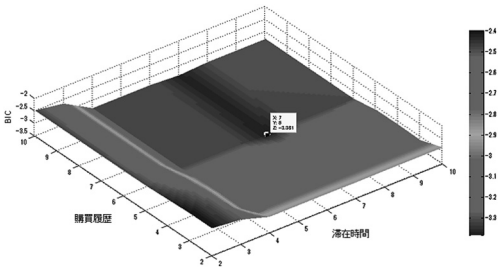

図 3 すべての組み合わせにおける BIC の密度推定

表 1 予測方法別の精度の比較

予測方法	的中率
線形判別分析 (LDA)	63.64%
ロジスティック回帰分析 (LR)	76.62%
ベイジアンネットワーク (BN)	77.49%

本実験ではモデルの検証方法としてホールドアウト法を使用しており，提案モデルと線形判別分析およびロジスティック回帰分析を比較している．線形判別分析およびロジスティック回帰分析は，プログラミング言語 R の「MASS」パッケージを利用した．われわれの提案モデルは上述した最適な組み合わせ，クラスター数 $L = 7$ および $M = 6$ を用い，図 2 および式 (4) において示されたネットワークが採用されている．表 1 はこれらの手法を比較した結果を示しており，表内の「的中率」は，テストデータにおける予測の精度を示している．そして提案モデルは，ほかの方法よりも高い精度を示している．

5.3 ROC 分析

分類問題では，判別値が 0 を下回っているか上回っているかによって，ターゲットは二つの異なったカテゴリへと区分される．確率モデルを適用した場合，基準値として 0.5 が広く使用されている．それにもかかわらず，陽性（陰性）データが少数グループに存在している場合には，結果において極端な歪度を引き起こす可能性がある．サンプルデータにおける的中率を最

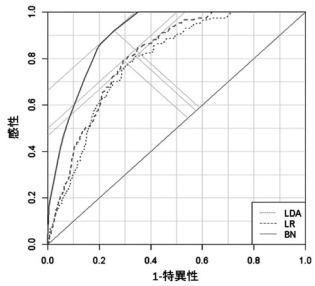

図 4　閾値の調整における ROC 曲線

表 2　AUC の比較

予測方法	AUC
線形判別分析 (LDA)	0.7883
ロジスティック回帰分析 (LR)	0.8094
ベイジアンネットワーク (BN)	0.9023

表 3　最適決定の閾値のもとでの複数の基準の比較

予測方法	閾値	感性	特異性	精度
LDA	0.2423	0.5217	0.4569	0.4762
LR	0.2713	0.7826	0.6235	0.6710
BN	0.3409	0.8841	0.7222	0.7706

大化するだけでは，タイプ I のエラー（タイプ II のエラー）を増加させてしまう可能性がある．それゆえに，モデルの性能は複数の基準に基づいて評価が行われる．その基準が感性および特異性であり，感性は陽性データに関する的中率，特異性は陰性データに関する的中率を示している．

図 4 の X 軸および Y 軸は，それぞれ 1 − 特異性（真陰性率）および感性（真陽性率）を示している．点線，破線，および実線は，それぞれ線形判別分析，ロジスティック回帰分析，および提案モデルの ROC 曲線を示している．表 2 において示されているように，提案モデルはほかの手法より高い精度を示している．また，表 3 は提案モデルのほうが，感性，特異性，および精度において，高い性能を有していることを示している．

5.4　技術的な諸課題

実際の分析において，説明変数の離散化は常に難しい問題の一つである．本稿では，よりよいクラスター数を求めるため，BIC を評価基準として採用した．しかしながら，BIC は限られたクラスター数の中で最適な解を求めることはできるが，さらに多様な説明変数が加われば，異なる離散化の方法論が必要となるであろう．

また上述したように，本稿で扱うような陽性・陰性データのバランスが極端な場合（いわゆる imbalanced data），分類クラスを決定する閾値の決定は単純ではない．最適な閾値の決定はモデルの分類精度に直接的な影響をもつため，今後，多様なケースに対応できる，最適な閾値を求める方法論の開発が求められる．

6.　おわりに

本稿は，RFID 技術を用いて収集された顧客の店舗内行動データについて，ベイジアンネットワークを用いた購買行動モデルを提案し，過去の購買傾向と売場での滞在時間を説明変数として採用している．本稿では説明変数の離散化などより実践的な分析プロセスを説明してきた．提案モデルは複数の観点からみても高い精度をもっており，実務に重要な知見を提供してくれるものである．今後，説明変数にブランドスウィッチに関連するような変数を導入し，特定の商品の販売促進にも有用な知見を抽出できるように，モデルを改良していくことを考えている．

謝辞　本研究はJSPS科研費基盤研究(A)16H02034の助成を受けたものです．

参考文献

[1] Y. Zuo and K. Yada, “Application of Bayesian network sheds light on purchase decision process basing on RFID technology,” In *Proceedings of 2013 IEEE 13th ICDM*, pp. 242–249, 2013.
[2] G. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, **6**(2), pp. 461–464, 1985.
[3] J. P. Egan, *Signal Detection Theory and ROC Analysis*, (*Academic Press Series in Cognition and Perception*), Academic Press, 1975.
[4] N. Havas and H. M. Smith, “Customers’ shopping patterns in retail food stores: An exploratory study,” U.S. Department of Agriculture, Agricultural Marketing Service, Marketing Development Research Division, Vol. AMS-400, 1960.
[5] W. D. Wells and L. A. Lo Sciuto, “Direct observation of purchasing behavior,” *Journal of Marketing Research*, **3**, pp. 227–233, 1966.
[6] H. Sorensen, “The science of shopping,” *Marketing Research*, **15**, pp. 30–35, 2003.
[7] J. S. Larson, E. T. Bradlow and P. S. Fader, “An exploratory look at supermarket shopping paths,” *International Journal of Research in Marketing*, **22**, pp. 395–414, 2005.
[8] S. K. Hui, E. T. Bradlow and P. S. Fader, “Testing behavioral hypotheses using an integrated model of grocery store shopping path and purchase behavior,” *Journal of Consumer Research*, **36**(3), pp. 478–493, 2009.
[9] K. Yada, “String analysis technique for shopping path in a supermarket,” *Journal of Intelligent Information Systems*, **36**, pp. 385–402, 2011.

