POS データからの売上変動パターン発掘

Sales Fluctuation Pattern Mining from POS Data

森本 康彦♥ 八木 一光♠

Yasuhiko MORIMOTO Kazumitsu YAGI

ある商品の売上の増減は，他の商品の売上の増減に影響を与えるという仮説にもとづき，多数の商品の日々の売上が記録されている POS データから商品の売上増減事象に着目した時系列パターン発掘を行った.本論文では，時系列パターンおよび推移的時系列パターンの確信度の最適化問題と，その空間効率のよいアルゴリズムについて簡単に述べ，実際の POS データから得られた解析結果と本手法の有用性について報告する.

We assume that sales fluctuation of an item must affect sales of another item. In this paper, we focus on fluctuations of sales of each item and find sequential patterns of the sales fluctuation events. In addition, we consider an optimization problem of transitive sequential pattern of the sales fluctuation. We developed and applied space efficient algorithms for computing optimized sequential patterns that maximize transitive confidence for large POS data and show its effectiveness.

1. はじめに

POS システムを導入すると，どの商品が，どの日時に,どの店舗で売れたか会計時に瞬時に記録することができる．日々の売上の集計や会計処理の効率化のため，現在までに，多くの小売業で POS システムが普及している．近年，POS システムに集計される情報は単なる会計業務の効率化の目的以外にも経営戦略，販売戦略をたてる上で重要な基礎情報として有効利用されている．例えば，店舗毎の売れ筋商品の見極め，商品，あるいは，商品カテゴリ毎の売上変動の傾向分析などが広く行われている．我々は，このような POS データのさらなる高度利用を目指し，POS データの分析に向くデータマイニング手法を開発し，その有効性を検証した．

分析対象のデータは平成 16 年度データ解析コンペティション[1]に提供された全国の複数の食料品店のもので,このデータ内には，JAN コード（バーコードに記される商品を識別するためのコード）単位で約 7 万品目の商品数があり，これらは JICFS 分類（日本で広く，かつ，標準的に利用されている加工食品，飲料，日用雑貨等をカバーする小売商品の分類）で約 200 種類のカテゴリにまたがっている．データは 2004 年 1 月から 6 月までの半年間の店舗および日付毎の商品別販売数量，販売金額の集計データである．また,店舗および日付毎の客数も関連データとして提供された．近年は，POS データに個人のポイントカードなどの FSP（Frequent Shopper Program）情報も含まれるケースがあるが，このような集計済みの形態だと，個人のプライバシは守られるので，今後，POS データの分析を外部に委託する場合はこのような集計済みのデータとなる場合が増えると想定される．

2. 時系列売上変動パターン

今回の分析では「ある商品の売上の増減は，他の商品の売上に影響を与える」という仮説をたて，商品および商品カテゴリの売上の変動事象の相関関係に着目し「$A^{+}\Rightarrow_{(i)}B^{-}$」のような形式を持つルールの発掘を行った．ルールの矢印の左辺と右辺の A,B は商品名，商品名に付与されている＋，－はその商品の売上の増減を，そして矢印の(i)は左辺と右辺の事象の時間間隔をそれぞれ表している．このルールは「商品 A の売上が上がった i 週間後に商品 B の売上が下がる」ということを示している．このような形式のルールは時系列パターンと呼ばれている[1,3,4]．時系列パターンは支持度(support)，確信度(confidence)という２つの指標で評価される．時系列パターンの支持度は左辺と右辺の両事象が時間間隔 i でこの順に発生する確率, つまり, このパターンの「適応度」を表す．一方，確信度は左辺事象が発生したという条件の下で，その時間 i 後に右辺事象が発生する条件付確率を示し，このパターンの「強さ」を表す．一般的に，データマイニングにおける時系列パターン発掘問題とはユーザの指定する最小支持度，最小確信度以上の時系列パターンを全て列挙する問題となる．

従来の時系列パターンでは，ある商品（あるいは商品集合を全て）を買ったか，買わなかったかという離散的な事象がパターンの左辺値，右辺値になるが，今回の分析では，ある商品の売上増減の変動が右辺値，左辺値となる．このような変動が起こったかどうかを，店舗および日付毎の商品別販売金額の集計値をもとに判定する．買ったか，買わないかといった離散的な事象と異なり，売上の変動は連続値であるため,この問題に既存のアルゴリズムあるいはマイニングツールを簡単に適用できない.

まず，食料品店の売上は曜日による季節変動が大きいため,その影響を受けないよう 1 週間毎に売上を再集計した. また,商品毎に単価は異なり，商品によっては同じ商品でも販売地域や店舗毎に販売価格が異なる．さらには，店舗の規模によって売上金額は大きく異なる．そのため，1 週間毎の店舗および日付毎の売上金額の集計値から客単価(売上金額/客数)を計算し，その増減幅を使って変動の判断をおこなう．商品によって客単価の日々の変動幅が大きいものや小さいものがありその変動トレンドも異なるため，商品毎に求めた客単価の変動幅をｚスコアで標準化する．ある商品のｚスコアが０なら売上変動傾向が変わらなかったことを表し，プラスなら変動傾向が売上増方向に変化した，マイナスなら売上減方向に変化したことを表す．ｚスコアの絶対値はその変化の大きさを意味し，この値が小さいときは売上変動傾向の変化が小さかったことになる．ユーザがｚスコアの閾値αを定め，その値によって増減の事象を判定する．たとえば閾値αを 0.5 とすると，ｚスコアが 0.5 増加（減少）した場合は変動傾向が増加（減少）方向に変化したとみなし，ｚスコアの変

♥ 正会員 広島大学総合科学部
morimoto@mis.hiroshima-u.ac.jp
♠ 非会員 広島大学大学院工学研究科情報工学専攻
yagi_hiroshima@hotmail.com

[1]経営科学系研究部会連合協議会共催のデータ分析技術，マイニング技術を競うコンペティション．毎年開催され本学会も共催している．

動の絶対値が0.5以下のときは売上傾向に変化がなかったとみなす.

時系列売上変動パターン発掘結果

売上増減のような連続事象であること，および，１商品に対し,買ったか,買わないかといった2状態ではなく,増加,減少，増減なしの3状態であることから，大規模なPOSデータベースに対して時系列売上変動パターンを発掘するためには,計算効率の面で,従来の問題以上の工夫が必要になる.我々は時系列売上変動パターンの効率的なアルゴリズムを開発し，実際のPOSデータでその有効性を確かめた．本論文では，分析結果と，この分析の有用性の速報としての報告にとどめ，アルゴリズムの詳細はスペースの都合上省略する.

マイニングに先立って最小支持度，最小確信度，増減閾値α，時間間隔の４つのパラメータをユーザが指定する.（ただし,我々のシステムでは，実装の都合上，一般的に使われる最小支持度の代わりに最小左辺支持度，つまり，左辺事象の発生確率の最小値を使った.「左辺支持度*確信度＝支持度」と読み替えることができる.）データに応じてこの４つのパラメータは適切な値が異なるため，様々な値の組み合わせを使って実行した．このデータに関しては，形態や規模,および，客層の類似する店舗が多かったためか，同じ売上変動パターンが各店で同様に起こっているというケースが多かった.そのため，全体的に支持度の高いパターンが多く存在していた.

試したパラメータのうち，最小左辺支持度=0.25，最小確信度=0.7，増減閾値α=1.0，時間間隔=１（時間間隔を１と短く設定したのは次章での推移的パターンで利用するためである．商品数が多いため，発掘されるパターンも多いが,論文として報告するパターンを重要なものだけに絞るという意図から，それ以外はやや高い値に設定した.）とした場合は,48,878個のパターンが発掘された.発掘されたこれらのパターンの概要の一部とそこから得られる知見（仮説）は以下のとおりである．尚，実際の結果では，左辺値，右辺値には，詳細な商品名（会社名，商品名，ブランド名）が入るが，データ提供元との契約で，商品を識別できる情報は伏せてある.

（特徴１）売上の上がった商品は，翌週，売上が減少する

この特徴に合致する例のうち確信度の高い上位３パターンは以下のようなものであった.

(口紅ピンクA)$^{+}\Rightarrow_{(1)}$(口紅ピンクA)$^{-}$
支持度0.28，確信度1.0

(健康茶ティーバッグB)$^{+}\Rightarrow_{(1)}$(健康茶ティーバッグB)$^{-}$
支持度0.32，確信度0.875

(有機濃厚ソースC)$^{+}\Rightarrow_{(1)}$(有機濃厚ソースC)$^{-}$
支持度0.28，確信度0.857

このようなルールは全48,878ルールのうち249個発掘された．反対に,「売上が下がった商品が翌週に売上が上がる」というパターンは8個しか発掘されなかった.このことから,商品の売上が上がると翌週の売上が下がりやすいが，１度下がった売上をすぐに上げるのは難しいと予測できる.

（特徴２）ドレッシング売上増の翌週，お茶の売上減

この特徴に合致する例には以下のようなものがあった.

(青じそドレッシングA)$^{+}\Rightarrow_{(1)}$(紅茶B)$^{-}$
支持度0.28，確信度0.857

(和風ドレッシングC)$^{+}\Rightarrow_{(1)}$(緑茶D)$^{-}$
支持度0.28，確信度0.71

このデータは１月から６月のデータであったが，寒い時期から暖かくなる過程で，顧客はサラダを食べる機会を増やし,（温かい）お茶を飲む機会を減らすようライフスタイルを変えていったのではないかと考えられる.

その他では「**ガムは左辺にも右辺にも現れない**」（ガムは計画性を持って購買されることが少ないと考えられる),「**マヨネーズは特定メーカの特定ブランドのみが左辺値に現れる**」(おそらく,この商品が目玉商品とされるケースが多い),「**右辺に現れる商品の多くは日用雑貨および化粧品である**」（これら２つのカテゴリの商品は他の商品の売上の影響を受けやすい．これらの商品を効果的かつ戦略的に販売することで収益性を上げることも可能であろう),「**左辺でプラスとして表れる商品の多くは嗜好飲料,日用雑貨,化粧品である**」（これらは目玉商品とされることが多いのであろう．日用雑貨と化粧品は，前述のように右辺値に来ることも多いが，それらが右辺にくる場合は，さらに別の商品の売上に影響を与えるので注意が必要であろう）といった特徴があった.

POSデータから得られる売上変動パターンを詳細に分析し,このような仮説としての知見を数多く得ることは有用で販売戦略を立てる上で重要である.

3. 推移的時系列パターン

時系列パターンでは左辺と右辺の２つの商品（あるいは商品集合）の相関関係しかわからず，場合によっては，これではより的確な分析ができない．例えば，前節で「右辺に現れる商品の多くは日用雑貨および化粧品」というパターンが多く見つかっていたが，同時に「左辺でプラスとして表れる商品の多くは嗜好飲料，日用雑貨，化粧品」というパターンも多い．これは「何かの事象⇒日用雑貨・化粧品$^{+}$⇒何かの事象」という推移的なパターンが多いことを示唆している.

時系列パターンの左辺，右辺の事象を頂点とすると,各パターンは左辺頂点から右辺頂点へ有向辺として表現することができる．また,パターンの確信度はその有向辺の重みと考えることができる．その場合，マイニングにより得られる時系列パターンの集合は巨大な重みつき有向グラフとなる.今，図１のような時系列パターン（図左）とその重みつき有向グラフ（図右）があるとする．図中のV1，…，V4の各頂点は売上変動の事象，矢印は時系列パターン，各パターンの数値はその確信度をそれぞれ表している.

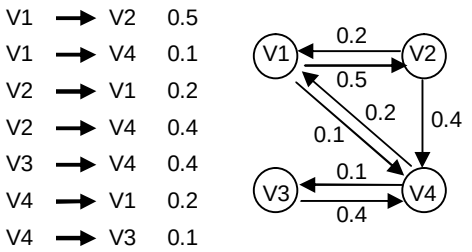

図1. 時系列パターンの有向グラフ
Fig.1 Weighted Directed Graph of Sequential Patterns

図１の「V1⇒V4」というパターンに注目する．このパターンだけを見ると確信度は0.1しかないため,V1の発生はそれほどV4に影響を与えないように思える．しかし，グラフをみると「V1⇒V2⇒V4」なる推移的な時系列パターンがありその推移的確信度（経路上の各確信度の積）は0.5*0.4=0.2と

なっており，もともとの確信度に比べるとかなり大きい値となっており，それは無視できない確率であるかもしれない．同様に，パターンの存在しない「V2⇒V3」，「V3⇒V1」などにも，ある程度の大きさの推移的な時系列パターンが存在していることがわかる．

左辺と右辺の２つの事象の時系列パターンのうち，前述の例のように左辺，右辺以外の事象を経由してよい時系列パターンを推移的時系列パターンとする．一般的にある２事象間の推移的時系列パターンは複数存在する．推移的確信度は各有向辺の（1 以下の値である）重みの積であるため，推移的時系列パターンの多くでは推移的確信度は非常に小さな値となる．推移的時系列パターンのうち，その推移的確信度がユーザの指定する最小確信度以上のものを「有効推移的時系列パターン」とし，ある２頂点間に存在する有効推移的時系列パターンのうち，その推移的確信度が最大なものを，その２頂点間の「最適推移的時系列パターン」（以降，これを簡単に「最適パターン」とよぶ）とする[5,6]．的確な販売戦略を考えるにはこのような最適パターンを知っておくことが重要である．

前章で述べた時系列売上変動パターンは確信度を重みとした重みつき有向グラフと考えることができる．最適パターンを求める問題は，このグラフ上に存在する頂点のすべてのペアのそれぞれの順列に対して，最適となる経路を求める問題とみなすことができ，この問題は全頂点間の最短経路アルゴリズムとして知られている「Floyd-Warshall 法」と呼ばれる動的計画法を応用して効率的に解くことができる[2]．しかし，この動的計画法の実行過程では，計算途中のその時点での全頂点間の（暫定）最適経路の情報をすべてメモリ中に保持しておかなくてはならず，この問題のような巨大なグラフを対象とした問題には利用できない場合も多い．

n 頂点からなるグラフの全頂点間の最短経路は，以下のような動的計画法で計算する．全頂点に 1 から n まで番号を与える．まず，全頂点間のどの頂点も経由しない経路（つまり，始点から終点への直接の有向辺）を「0時点での最短経路」とする．「1 番から k 番までの頂点のみを経由する」全頂点間の最短経路を「k時点の最短経路」とすると，k＋1 時点の最短経路における，i 番頂点から j 番頂点への最短経路は『k 時点の「i 番から j 番への経路」』または『k時点の「i 番から k 番への経路」と「k 番から j 番への経路」を繋いだ経路』のいずれかとなる．この事実に基づき k=0 から n まで「k 時点の最短経路」を計算すると最短経路が求まる[2]．この動的計画法を実行するためには「k 時点の最短経路」をメモリに持っておく必要があるが，これを図２のような CE-ハッシュと呼ぶデータ構造で保持する．図２は図１のグラフの k=4 の「k 時点の最短経路」の情報である．例えば，V1 から V4 へ経路を考える．図の左のエントリは，始点に対応するハッシュエントリとなっており，そのエントリの V1 を調べる．V1 のエントリにはさらに終点のハッシュエントリがあり，V4 に対応するエントリ内に V1 から V4 への経路情報である C_{14} と記されたセルが見つかる．セルの点線は接頭経路を示しており，点線がなくなるまで辿ると（終点から始点へ逆向きの）経路がわかる．この場合 V4←V2←V1 が最短経路でその推移的確信度は 0.2 となっている．このデータ構造で「k時点の最短経路」の情報が効率的に保持できる．さらに，動的計画法の計算を進める過程で，その後の計算で使う必要がない経路情報（セル）を効果的に CE-ハッシュから枝狩りするアルゴリズムを実装し，巨大なグラフに対しても最適パターンを実行できるようにした[5,6]．

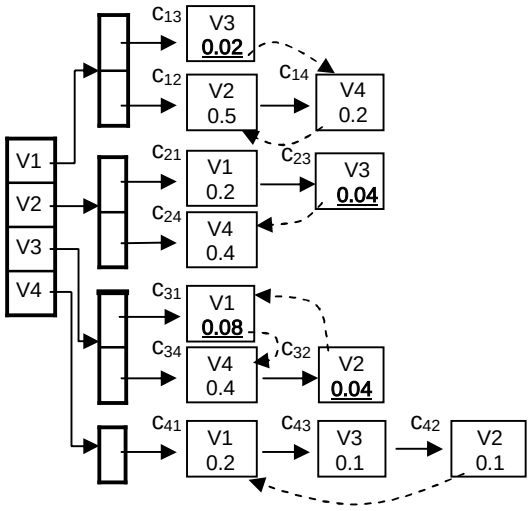

図 2. CE-ハッシュ
Fig.2 CE-Hash

最適パターン発掘結果

今回の POS データからは，前章の分析結果として 48,877 個の時系列売上変動パターンが得られた．これらのパターンからなるグラフには 12,941 個の頂点（商品売上の変動事象）と 48,877 個の有向辺があった．このグラフから最小確信度を 0.7 として最適パターンを計算すると 51,979 個の最適パターンが発掘された．そのうち，左辺，右辺以外の事象を経由している最適パターンは 3,105 個であった．このうち経由する事象が１つであったのは 2,991 個，２つの事象を経由するものは 114 個であった．今回のデータでは３つ以上の事象を経由するものは最適パターンとならなかった．これらのパターンは従来の時系列パターンマイニングで見つかるパターン（つまりグラフにおける直接の経路）より他の事象を経由した推移的なパターンの確信度のほうが大きくなっている．いうまでもなく，こちらの，より確信度の大きい値のほうが適切な販売戦略を立てる上では重要であるため，これらを多数発見できたことの意義は大きい．左辺から右辺への確信度が，他の事象を経由したほうがよくなるケースで，その確信度の増分が最大であったものは以下のパターンであった．（増分最大のパターンは複数あったがそのうち一つだけを例示する．）

(口紅 A)⁻ ⇒(1) (オレンジ口紅 B)⁺
確信度 0.0

(口紅 A)⁻ ⇒(1) (綿棒パック C)⁺ ⇒(1) (オレンジ口紅 B)⁺
確信度 0.857

「口紅 A」の売上減と「オレンジ口紅 B」の売上増という時系列パターンは存在していなかったため，一見するとこの２つの事象には相関がないように思えるが「綿棒パック C」の売上増がそれらの間に起こった場合には推移的確信度の高い時系列パターンが存在していることがわかる．

今回，発見された最適パターン全体の特徴を把握するため，最適パターンの経由事象（つまり，最適パターンの左辺でも右辺でもない事象）に関する以下の表１，表２のような統計を調べた．表１は経由事象として最適パターンに出現する回数の多い上位５商品である．最適パターンが約３千個なので，

これらはいずれもその 2%以上をしめる．上位のほとんどが化粧品の売上減である点が興味深い．また，表２は経由事象として最適パターンに現れる回数を商品カテゴリのプラスおよびマイナス毎に集計したものである．やはり，化粧品のマイナスが多いことがわかる．また，化粧品，日用雑貨，嗜好飲料は経由事象として表れる場合が多いことが改めて確認された．とくにこれらの売上の増減について考える場合は，その前後の因果関係に注意する必要があるであろう．また，経由事象は売上減として現れる傾向が高いことがわかる．

表１．最適パターンの経由事象としての出現数（上位５位）
Table 1. Number of Appearance as Transitive Events in Optimized Sequential Patterns (Top 5)

売上変動(経由)事象	出現数
アイブロウペン A(-)	84
口紅 B(-)	79
口紅ピンク C(-)	75
アロマオイル D(-)	74
口紅 E(-)	74

表２．最適パターンの経由事象としてのカテゴリ別出現数
Table 2. Number of Appearance by Category as Transitive Events in Optimized Sequential Patterns

カテゴリ	(+)数	(−)数
化粧品	805	1278
日用雑貨	277	313
嗜好飲料	82	143
砂糖	15	78
ソース	0	20
味噌	13	18
ドレッシング	5	16
醤油	18	14
キャンディ	39	14
クッキー	2	12
マヨネーズ	0	12
米菓	8	11
スナック	2	7
チョコレート	16	0

4. まとめ

本論文では，食料品店の大規模 POS データからの売上変動パターンの発掘手法とその有効性について検証した．データマイニング技術の多くは，併売の組み合わせとなる頻出パターンを見つけるマーケットバスケット分析を想定しており，この分野での応用範囲は広いと考えられていた．併売の傾向は商品の棚割りなどに実際に応用されているが，近未来の仕入れの戦略や販売戦略を考える際には，それだけでは不十分で順序を考慮した時系列パターンを知る必要性が高まる．既存の時系列パターンマイニングは，その期待に応えうるものであるが，それには個人を特定できるデータが不可欠である．近年は個人情報保護の考えが浸透しているためそれらの情報を利用しないデータマイニングも必要となるであろう．今回は，店舗および日付毎に集計された商品の販売数量，販売金額といった「個」ではない「マス」情報から，時系列売上変動パターン，最適推移的時系列売上変動パターンを発掘し，このような集計情報からでも，販売戦略に有効な知見を得ることができた．

集計情報からのマイニングは「売上金額」などのような連続値事象に対するマイニングであるため既存の離散事象を対象とする一般的なマイニングツールではうまく処理できない．既存のマイニングツールでも扱えるような前処理を施せばよい場合もあるが，このような集計データからマイニングするというケースも増えるため，あらかじめ必要な前処理部分も含めた効率化問題も今後重要となるであろう．今回，我々が行った手法でも，全体としての効率化をおこなったが，この面では改善の余地はまだ大きいと考える．

本論文の３章で述べた最適パターンは，マイニング手法としては新しいものである．今回の売上変動パターンでもその有効性の一端を確認できたが，これはもちろん個人を識別できる情報を利用できる場合には，個人の購買行動の細かい分析に応用できるであろう．近年，CRM（Customer Relationship Management）の重要性が広く認識され，ポイントカード，マイレージカードなど FSP が広まっているので，そうした FSP 情報が利用できる環境では様々な応用が期待できる．

[謝辞]

平成 16 年度データ解析コンペティション事務局の皆様，同コンペティションに貴重な実データを提供してくださった匿名の企業様，また，これまで本研究・開発に従事してくださった空本麻衣，茂久田美穂の各氏に深く感謝いたします．

[文献]

[1] R. Agrawal, R. Srikant, “Mining Sequential Patterns,” Proc. of the IEEE ICDE, pp. 3-14, 1995.

[2] T. Cormen, C. Leiserson, R. Rivest, C. Stein, “Introduction to Algorithms,” MIT Press, 2001.

[3] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, M. Hsu, “FreeSpan: Frequent Pattern-projected Sequential Pattern Mining,” Proc. of the ACM SIGKDD, pp. 355-359, 2000.

[4] H. Mannila, H. Toivonen, A. Verkamo, “Discovery of Frequent Episodes in Event Sequences,” Data Mining and Knowledge Discovery, Vol. 3, pp. 259-289, 1997.

[5] Y. Morimoto, “Optimized Transitive Association Rule: Mining Significant Stopover between Events,” Proc. of the ACM SAC, pp.547-548, 2005.

[6] K. Yagi, M. Soramoto, M. Mokuda, Y. Morimoto, “Optimized Sequential Pattern Mining from Point Of Sales Data,” Proc. of the IEEE Int’l Special Workshop on Databases for Next Generation Researchers, pp. 12-15, 2005.

森本 康彦 Yasuhiko MORIMOTO
広島大学総合科学部助教授．1991 年 広島大学大学院工学研究科博士課程前期修了．同年 日本 IBM㈱東京基礎研究所 2002 年 12 月より現職．博士（工学）．データマイニング，GIS，オブジェクト指向データベース等の研究・開発に従事．日本データベース学会，情報処理学会，日本ソフトウエア科学会，ACM，IEEE CS 各会員．

八木 一光 Kazumitsu YAGI
広島大学大学院工学研究科情報工学専攻博士課程前期 在学中．リンク分析，時系列相関ルールなどデータマイニングアルゴリズムの研究・開発に従事．