2J4-OS-04a-1

ACO型時系列パターン抽出法を用いたマーケティングデータの考察

Consideration of marketing data using the ACO-based pattern mining method

坪井一晃 *1 篠田孝祐 *1 諏訪博彦 *2 栗原聡 *1
Kazuaki Tsuboi Kosuke Shinoda Hirohiko Suwa Satoshi Kurihara

*1電気通信大学大学院情報システム学研究科
Graduate School of Information System, The University of Electro-Communications

*2奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

It is important to understand a pattern of consumer needs correctly and clarify target of goods and service in marketing. It is considered that consumer action always changes. Then, we focus on the ACO algorithm having high robustness and adaptability. We analyze the real marketing data by using sequential pattern mining based on the ACO algorithm. In this paper, the real marketing data is visualized the relation of goods that they are brought with by consumer.The data was offered by Joint Association Study Group of Management Science.

1. はじめに

マーケティングを考える上で，多様な消費者のニーズを正しく理解し，的を絞った商品やサービスが求められている．また，サービスを展開するためには，消費者の行動や心理といった消費者インサイトを理解することは重要である．現実における消費者の行動を考えてみると，季節の変化や世間の流行の変化によって，人の行動も変化していることが容易に想像できる．よって，消費者の行動の分析においても，変化を受容できる柔軟なシステムが必要とされる．

そこで，我々は，Ant Colony Optimization（以下，ACO）アルゴリズムを応用したパターンマイニング技術の構築を目指している．ACO アルゴリズムとは，アリの採餌行動をモデル化した最適化手法として知られている．アリは，現実世界において協調することによって，アリの巣と餌場における最短経路を得ることができる．無論，餌場における餌がなくなった場合には新たに餌の探索を行い，新しい餌場を見つけると新たな餌場との最短経路の取得を行う．この様子を巡回セールスマン問題における最適化モデルとして提案されたのが ACO アルゴリズムである．本研究では，ACO アルゴリズムがもつ高い頑健性や適応性に着目し，パターンマイニングに応用することを試みる．

本論文の構成を次に示す．2 章では，パターンマイニングの研究例や，本研究で着目した ACO アルゴリズムに関する研究例について述べる．3 章では，解析手法となる ACO 型時系列パターン抽出アルゴリズムについて述べる．4 章では，提案手法を用いて，北海道江別市にある一店舗における実データを解析した結果を示す．5 章で，北海道江別市にある二店舗と北海道札幌市にある一店舗を一つのグラフにまとめて可視化することで店舗間の差異について考察する．6 章で，まとめと今後の課題について述べる．

連絡先: 坪井一晃，電気通信大学大学院情報システム学研究科，東京都調布市調布ヶ丘 1 － 5 － 1，TEL042 － 443 － 5664，mail:tsuboi@uec.ac.jp

2. 関連研究

頻出するパターンを発見するアルゴリズムとしてアプリオリアルゴリズム [1] が有名である．他にも，データベース上から頻出パターンを発見するためのアルゴリズムとして，時系列を考慮したパターンを発見する Apriori All[2] をはじめ，GSP アルゴリズム [3]，SPADE アルゴリズム [4] などが提案されている．しかし，これらのアルゴリズムでは，全ての入力データを等しく参照しているために，変化する消費者ニーズに対応した頻出パターンの抽出は達成できない．

そこで，我々は ACO アルゴリズムに着目する．ACO アルゴリズムでは，アリは通過した経路にフェロモンを残すことと，フェロモンの濃度が濃い経路を好んで経路選択をすることという行動を前提として考えている．この前提によって，アリが集団として行動するごとに，最短経路を通過するアリが多くなり，最短経路に残るフェロモンの濃度が濃くなる．一方，フェロモンは気体であり，時間経過に伴い蒸発する．アリが経過する頻度が少ない経路について，フェロモンの濃度が薄くなる．残ったフェロモンの濃度が濃い経路が最短経路の解となる．ACO アルゴリズムの特徴として，高い頑健性および適応性を有することが挙げられる．

ACO アルゴリズムをパターンマイニングに応用した事例として玉置ら [5] の研究が挙げられる．玉置らはセンサが人の行動を読み取ることに着目し，連続した人の行動からセンサの隣接関係を，ACO アルゴリズムを用いて推定する研究を行っている．ACO アルゴリズムを用いたことで，センサの故障や移動といった環境の変化などのノイズにも対応できる，センサの隣接関係の推定を達成した．このように，もともと最適化技術で知られていた ACO アルゴリズムであるが，単純な行動ルールに基づくエージェントの移動と環境に対するフェロモンの付加・蒸発を応用することで最適化以外の分野でも活用できる．本研究においても ACO アルゴリズムが有する動的安定性や頑健性に着目し，ACO 型時系列パターン抽出アルゴリズムを用いて対象データの解析を試みる．

3. ACO に基づくパターン抽出手法

本研究では，小売店における消費者の購買結果であるレシート情報から，頻出する購買パターンを抽出する．本研究における頻出する購買パターンとは，同時に購入される頻度が高い組み合わせとする．

3.1 入力データ

提案アルゴリズムを適用するにあたって，入力データセットは消費者の購買結果であるレシート情報である．一枚のレシートに記載された商品名に着目し，アルゴリズムを適用する．また，レシートには購入された時刻が記載されているため，この時系列順にレシート情報を逐次的に入力していく（図 1）．$c_{t,k}$ はレシートを表し，添え字 t は購入年月日を表し，添え字 k は同日内のレシートの番号を表す．なお，レシートの番号は時系列順に連番が割り振られている．s_i は商品名を表す．図 1 を例にすると，一番初めのレシート $c_{20140101,1}$ では，商品 s_1, s_2, s_3, s_4 が購入されおり，次のレシート $c_{20140101,2}$ では，s_5, s_6, s_7 が購入されているといったような，レシートに記載される商品名を，レシートの時系列に沿って並べている．ただし，一枚のレシートの中に現れた同一商品の重複は無視することにした．

$$
\begin{aligned}
c_{20140101,1} &= \{s_1, s_2, s_3, s_4\} \\
c_{20140101,2} &= \{s_5, s_6, s_7\} \\
c_{20140101,3} &= \{s_3, s_8, s_9, s_{10}\} \\
&\vdots \\
c_{20140102,1} &= \{s_5, s_{11}\} \\
c_{20140102,2} &= \{s_{12}\} \\
&\vdots \\
&\vdots
\end{aligned}
$$

図 1: 入力データ

3.2 ACO に基づくパターン抽出アルゴリズム

ACO アルゴリズムは環境へのフェロモンの付加フェーズと蒸発フェーズから成り立つ．一日分のレシート情報を読み込み環境へフェロモンの付加フェーズを実行する．その後，環境に残るフェロモンに対して蒸発フェーズを実行し，次のフェロモン付加フェーズに移る．

まず，環境にフェロモンを付加するために，仮想グラフ G を用意する．$G = (V, E)$ は仮想空間上の無向グラフである．グラフ G における各ノード $v_i \in V$ は実環境における各商品 $s_i \in S$ に対応し，各エッジ $e_{i,j} \in E$ は商品 s_i と商品 s_j を同一レシート内に出現したことを表す．また，グラフ G 上に付加されるフェロモンの分布は τ で表す．なお，前提条件として出現する購買パターンに対する予備知識がないため τ の初期値は 0 とした．

次に，フェロモンの付加フェーズについて述べる．レシート c ごとにレシート内に現れた商品同士の度数分布 $d_{i,j}(c_{t,k})$ を作成する．度数分布 $d_{i,j}(c_{t,k})$ はレシート内に商品 s_i と商品 s_j が同時に出現した回数を表す．一日分のレシート情報から度数分布 $d_{i,j}(c_{t,k})$ を作成し，式 1 にしたがって，フェロモン τ を更新する．なお，t は日数を表す．

$$\tau_{i,j}(t) = \tau_{i,j}(t-1) + \sum_k d_{i,j}(c_{t-1,k}) \tag{1}$$

フェロモンの蒸発フェーズを実行した後，フェロモンの蒸発フェーズに移る．フェロモンの蒸発フェーズでは，式 2 に従い，環境に残るフェロモンを減少させる．なお，蒸発は蒸発率 ρ に依存する．

$$\tau_{i,j}(t) = \tau_{i,j} \times (1 - \rho) \tag{2}$$

このフェロモン蒸発フェーズを実行することにより，古い情報は少しずつ破棄され，解析結果に常に新しい情報が一定の割合で反映されることになる．ここで，蒸発率 ρ は情報の更新の速さを表す．

3.3 出力結果の可視化

出力となる仮想グラフ G は cytoscape を利用して可視化する．可視化させる際に，仮想グラフ G に対して，残量フェロモン量の閾値 ϕ を決め，閾値 ϕ に満たないエッジを削除する．次に，残された仮想グラフ G に対して Cytoscape のプラグインである cluster maker2 の community clustering を実行する．この community clustering は Girvan-Newman アルゴリズムによるクラスタリングが実装されたものである．クラスタリングによって得られた結果を利用して，各クラスタをそれぞれ円形にならべる．それぞれの円の位置は，クラスタ間のエッジが極力みえるように試行錯誤の上，配置する．

4. 実データを用いた実験

ACO に基づくパターン抽出アルゴリズムによって，経営科学系研究部会連合協議会主催平成 26 年度データ解析コンペティションで提供された実際のマーケティングデータの解析を行う．実験では，全日食チェーンの北海道江別市の店舗における 2013 年 7 月 1 日から 2014 年 6 月 31 日までのレシートデータを用いる．レシート数は 115756 レシート，登場する商品名数は 7082 商品，一日あたりの平均レシート数は 318（最大レシート数は 425，最小レシート数は 132）であった．蒸発率 ρ は 0.01 に設定した．

2013 年 12 月 31 日分までのレシートデータを入力したときの結果（図 2）と 2014 年 6 月 30 日分までのレシートデータを入力したときの結果（図 3）を掲載する．なお，エッジを削除するための閾値は $\phi = 25$（図 2），$\phi = 28$ （図 3）とした．閾値の設定は，実験結果を考察しやすくすることを考慮して決定した．また，各クラスタを区別するため，各クラスタにアルファベットを順に振り当てて表記した．

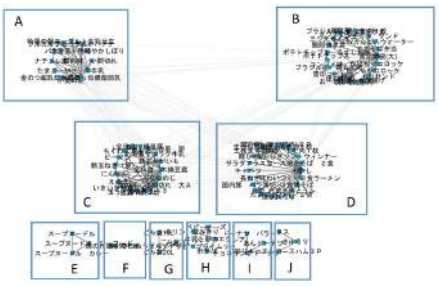

図 2: 2013 年 12 月 31 日分までのレシートデータを入力したときの出力結果

図 2 におけるクラスタ A をみるとヨーグルトや牛乳，たまご，納豆，豚肉が並んでいる．豚肉を除いて考えると朝食のための食品と考えられる．このように，実験結果の図は，消費者の購買パターンを可視化した結果といえる．

図 3 におけるクラスタ B を見ると，やきそば弁当，鶏照焼き丼，かぼちゃコロッケ，肉じゃがコロッケ，カツサンド，唐揚げ，さつまいも天，メンチカツサンド，飲料，ポテトチップス，買物袋，鶏肉が並んでいる．昼ごはんや軽食用に購入するパターンのクラスタと考えられる．しかし，図 3 を見ると先に述べた昼ごはんや軽食用の購入パターンと考えられるクラスタが存在しない．代わりに，クラスタ L では，お茶が，クラスタ M ではポテトチップスが独立したクラスタとして存在

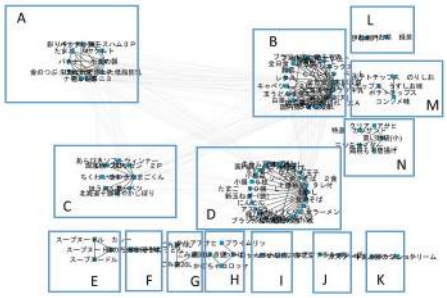

図 3: 2014 年 6 月 30 日分までのレシートデータを入力したときの出力結果

するようになった．また，クラスタ M をみると唐揚げ，カツサンド，サイダーと買物袋のクラスタが抽出されている．そして，焼きそば弁当や鶏照焼き丼といった商品が図から消えている．このことから，12 月までのレシートデータを入力したときには昼食やおやつ，軽食といった抽象度が高い購買パターンであったのに対して，6 月までのレシートデータを入力したときには，お茶やポテトチップスといった具体的な目的商品を購入するパターンが増えると考察できる．

5. 店舗間の比較

消費者の購買パターンを店舗間で比較することで，店舗がもつ特徴を調べる．今回の実験における比較対象とした店舗は次の三店舗である．また，各店舗を区別するために，店舗 A，店舗 B，店舗 C と表記する．

- 店舗 A：北海道江別市
- 店舗 B：北海道札幌市
- 店舗 C：北海道江別市

5.1 各店舗におけるグラフの生成

まず，3 章で述べた ACO に基づくパターン抽出法に基づいて，それぞれの店舗についての入力データセットからフェロモンの付加と蒸発を行った仮想グラフを生成する．ここで，各店舗における頻出する消費者の購買パターンを比較するという目的から，各店舗における仮想グラフ上で，残留するフェロモン量 $\tau_{i,j}$ の濃い順にノード数が 200 となるように仮想グラフ上のノードを削除する．

5.2 商品名の修正

店舗間の比較を行う際に，店舗によって商品名の取り扱いが異なることがあることを考慮し，商品名が異なることが多い野菜や肉を主に商品名の調整を行った．なお，個数や内容量に差がある場合に関しても同一商品名になるように調整した．調整の具体例としては,「きゅうり」や「きゅうり　2 本」,「キュウリ　2 本」,「キュウリ　3 本」,「胡瓜（バラ）」を「キュウリ」とすることや,「国内豚　バラ薄切り」や「国内豚　バラ薄切り（小）」,「国内豚　バラうすぎり（大）」を「豚バラ肉」とすること,「コカ・コーラ」や「コカコーラ　OTG ボトル」,「コカコーラ　コカコーラ P」を「コカコーラ」とした．

5.3 グラフの合併

それぞれの店舗についての頻出する消費者の購買パターンを表す仮想グラフを合併させる．仮想グラフを合併させる際に，店舗による差異が確認しやすいように，店舗の差を色で表現する．具体的には次の通りである．

- 店舗 A においてのみで抽出された頻出パターンを表すエッジは赤色
- 店舗 B においてのみで抽出された頻出パターンを表すエッジは緑色
- 店舗 C においてのみで抽出された頻出パターンを表すエッジは青色
- 店舗 A と店舗 B において抽出された頻出パターンを表すエッジは黄色（赤色＋緑色）
- 店舗 A と店舗 C において抽出された頻出パターンを表すエッジは桃色（赤色＋青色）
- 店舗 B と店舗 C において抽出された頻出パターンを表すエッジは水色（緑色＋青色）
- 店舗 A と店舗 B と店舗 C の全ての店舗において抽出された頻出パターンを表すエッジは黒色

それぞれの店舗について仮想グラフを合併させた仮想グラフに対しても，3.3 節で述べた可視化方法と同様に Cytoscape のプラグラインである Cluster Maker2 の community clustering を実行し，得られた各クラスタを円形に成形した後，それぞれの円を試行錯誤で配置した．

2013 年 12 月 31 日分までの半年分のレシートデータを入力したときの三店舗の比較結果が図 4 であり，2014 年 6 月 30 日分までの一年分のレシートデータを入力したときの三店舗の比較結果が図 5 である．

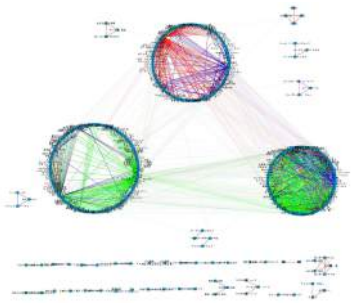

図 4: 2013 年 12 月 31 日分までのレシートデータを入力したときの出力結果

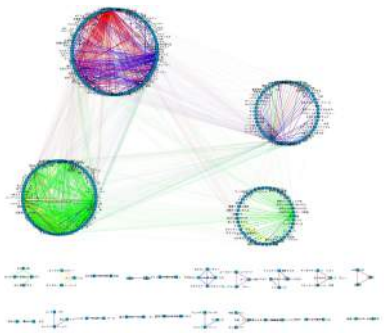

図 5: 2014 年 6 月 30 日分までのレシートデータを入力したときの出力結果

図 4 をみると，クラスタ A 内では，赤色と青色のエッジが多数を占めていることが確認できる．一方で，クラスタ C 内

では緑色のエッジが多数である．クラスタ C では，緑色が主であるが他の色も多く確認できる．つまり，クラスタ A が店舗 A と店舗 C の特徴を表し，クラスタ B が店舗 B の特徴を表すものと考えられる．クラスタ A に属している商品は，かっぱえびせんやポテトチップス，じゃがりこなどのおかし類，たこ焼きやフライドチキン，フランクフルト，メンチカツサンド，かぼちゃコロッケ，えびカツサンド，やきそば弁当，カツ丼，鶏照焼き丼などの弁当や惣菜モノ，緑のたぬきや赤いきつね，カリー屋カレーなどのレトルト食品，CC レモンやコカコーラ，オロナミン C やおーいお茶などの清涼飲料水，クリアアサヒやスーパードライ，淡麗生，ドラフトワンなどのお酒類，ラークマイルドやセブンスターといったたばこ類，買物袋などが並んでいる．

一方，クラスタ B をみると，トマトやとうもろこし，じゃがいも，ほうれん草，れんこん，チンゲン菜，ナスなどの野菜類，生ラムや豚ロース，豚バラ肉などの肉類，さんまやさば，えび，マグロやイカ，ほっけ，サーモンなどの魚類，スプライトやヤクルト，ソフトカツゲン，アイスコーヒーなどの清涼飲料水，ポテトサラダやマカロニサラダ，いなり寿司などの惣菜，みんなの食パンアンやバターロール，クルミパン，絹艶，菓子パンなどのパン類，みかんやバナナといった果物，からしやしょうが，わさびといった調味料，冷凍餃子や 3 食ラーメン，やきそば，メンマなどの簡易料理およびそれに付随するものやゴミ袋が並んでいる．

2013 年 12 月 31 日分までのレシートデータを入力した結果において，店舗 A と店舗 C の店舗の特徴がクラスタ A に表れていると考えると，この二店舗はおかしや惣菜，弁当といった，その商品だけでも食べられるものや調理をするにしてもレトルト食品のような簡単なものを求めるような購買パターンが特徴であるといえる．また，一方で，店舗 B の特徴がクラスタ B に表れていると考えると，野菜や肉，魚といった食べるためには調理などの一手間加えなければならない商品の購入パターンが特徴であるといえる．

図 5 をみると，クラスタ A 内では赤色と青色のエッジが多数を占めていることが確認できる．一方でクラスタ C とクラスタ D では緑色のエッジが多数確認できる．先と同様に，クラスタ A が店舗 A と店舗 C の特徴を表し，クラスタ B とクラスタ D が店舗 B の特徴を表していると考えられる．クラスタ A に属している商品はかっぱえびせんやポテトチップス，スーパーカップ（アイス），プリン，ヨーグルト，ココナッツサブレ，クリームドーナツなどのおかし類，たこ焼きやフライドチキン，フランクフルト，メンチカツサンド，かぼちゃコロッケ，えびカツサンド，鶏照焼き丼，唐揚げ弁当などの弁当や惣菜モノ，CC レモンやコカコーラ，オロナミン C，おーいお茶などの清涼飲料水，クリアアサヒやスーパードライ，金麦などのお酒類，レジ袋，鶏モモ肉や豚小間などの肉類などが並んでいる．

一方，クラスタ B をみると，トマトやじゃがいも，ほうれん草，チンゲン菜，ナス，にら，ぶなしめじ，ごぼう，まいたけ，エリンギ，キュウリ，サツマイモ，ピーマン，かぼちゃ，長いも，長ネギなどの野菜類，豚バラ肉や豚モモ肉，豚ロースといった豚肉，リンゴやバナナ，グレープなどの果物類などが並んでいる．また，クラスタ D をみるとのり弁当，焼きそば弁当といった弁当，サーモンやほっけ開き，いか刺身といった魚，緑のたぬきや赤いきつね，どん兵衛といったカップ麺，淡麗や麦とホップ，黒ラベルといったお酒，ポテトチップスなどのおかし，買物袋が並んでいる．

2014 年 6 月 30 日分までの一年分のレシートデータを入力した結果において，基本的には，2013 年 12 月 31 日分までのレシートデータを入力した場合に似た結果といえる．しかし，変化している点も多々見受けられる．例えば，図 4 におけるクラスタ B は野菜，肉，魚や果物と幅広い範囲のものであったものが，図 5 におけるクラスタ B は野菜と果物が多数を占めるクラスタとして抽出されるようになった．これは 12 月から 6 月にかけて，店舗 B において，消費者の野菜と果物の枠組みのなかで購入する購買パターンが増加したと考えられる．また，図 5 におけるクラスタ D に着目すると，クラスタ内のエッジがある一点の商品，買物袋に集中していることがわかる．その結果，クラスタ D には，雑多にさまざまな商品が並んでいると考えられる．また，店舗による差異を考えると，店舗 A と店舗 C が似ており，出来た上がりの料理であったり，おかしなどの調理が不要な商品の購入が特徴であり，店舗 B は野菜や肉といった料理の材料の購入が特徴であると考えられる．

6. おわりに

本研究では，ACO に基づくパターン抽出法を用いてマーケティングデータの解析を行った．全日食チェーンの実際のマーケティングデータを解析し，2013 年 12 月 31 日分までの半年分のレシートデータを入力した際に得られる購買パターンと 2014 年 6 月 30 日分までのレシートデータを入力した際に得られる購買パターンを抽出し可視化することで，購買パターンの差異について考察した．また，店舗の特徴を調査するために，三店舗についてのマーケティングデータを解析し，店舗ごとに色分けし可視化した．店舗による色分けした結果を考察することで，三店舗のうち二店舗が似ている特長を有することがわかった．

今後の課題として，商品名を調整する際の粒度を再検討する必要がある．本稿において，著者の判断で野菜や肉についてを主に調整したが，例えば，肉に対する調整の粒度を考えてみると，牛・鶏・豚のみによる分類や国産・外国産といった産地に関する分類，モモやバラ，ヒレ，ロースといった部位に関する分類を行うか行わないかといった，どこまでを分類するかといった問題がある．また，ACO に基づくパターン抽出手法においても，蒸発率などのパラメータが存在するために，パラメータによる結果の変化も考察する必要がある．

また，頻出する商品を結合させることでノードの粒度を高め，アルゴリズムを階層的に扱うことでより抽象度の高い情報の取得にも試みたいと考える．

参考文献

[1] Agrawal, Rakesh, and Ramakrishnan Srikant. ”Fast algorithms for mining association rules.” Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215. 1994.

[2] Agrawal, Rakesh, and Ramakrishnan Srikant. ”Mining sequential patterns.”Data Engineering, 1995. Proceedings of the Eleventh International Conference on. IEEE, 1995.

[3] Srikant, Ramakrishnan, and Rakesh Agrawal. Mining sequential patterns: Generalizations and performance improvements. Springer Berlin Heidelberg, 1996.

[4] Zaki, Mohammed J. ”SPADE: An efficient algorithm for mining frequent sequences.” Machine learning 42.1-2 (2001): 31-60.

[5] Tamaki, Hiroshi, et al. ”Pheromone Approach to the Adaptive Discovery of Sensor-Network Topology.” Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 02. IEEE Computer Society, 2008.