4N2-OS-01b-4

群知能メカニズムによる時系列階層パタン構造抽出法の提案

The proposal of extraction of time series pattern with hierachical structure based on collective intelligence

坪井一晃 *1　須賀聖 *2　栗原聡 *2
Kazuaki Tsuboi　Satoshi Suga　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電気通信大学大学院情報理工学研究科
Graduate School of Informatics and Engineering, The University of Electro-Communications

The patterns of behavior in everyday situations such as home, office, and commuting, and how to sell goods by day of the week, season, location requirements, include patterns that have hierarchies of various temporal granularity. Normally, it is often analysis in such hierarchy or in many cases in advance giving the hierarchical structure. Therefore, in this research, we propose a method to automatically extract both hierarchical structure and pattern from time series data using collective intelligent method.

1. はじめに

インターネットの急激な進化やコンピュータの高性能化と低価格化を背景に，社会のいたるところに IT を活用したシステムが導入されるようになっている．導入されたシステムを利活用するたびに，システムからデータが生成されて続けている．我々の意図にかかわらず稼働し続けるシステムも存在し，日々多種多様にわたるデータが大量に生成・蓄積されている．近年，この蓄積されるデータを用いることで有用な情報を得られるのではないかという期待が高まっている．このような背景から，蓄積されたデータから有用な知見を得るための技術としてデータマイニングに注目が集まっている．

データマイニングにおける重要な技術の一つに，データベースの中から頻出するアイテムの組み合わせを抽出するパターンマイニングがある．パターンマイニングを用いることで，例えば，コンビニなどの小売店におけるよく一緒に購入される商品などを効率的に発見できるようになる．さらに，現実世界において度々重視される時系列に注目した時系列パターンマイニングの技術へと発展を見せている．時系列を考慮したパターンを抽出することによって，この商品が売れた次にはこの商品が売れるはずだといったパターンが抽出できるようになった．これらのパターンマイニングでは，解析対象のデータベースはトランザクションと呼ばれる単位で分割されている．したがって，このトランザクションの中という範囲のなかでパターンと考えられるアイテムの組み合わせを抽出している．つまり，データベース中から分析したい粒度を明確に設定し，それに合わせたトランザクションに分割する．複数の異なる粒度の分析を行うときには，それぞれの設定による分析をその都度行う．

しかし，現実世界を考えたときに，日常のおける自宅とオフィスなどをつなぐ通勤といった行動様式や小売店における曜日や季節などによる商品の売れ方などの中には，様々な時間的粒度のパターンが含まれているのは自明である．そこで，本研究では，異なる時間的粒度を階層的にとらえ，時系列データから階層構造とパターンの両方を自動的に抽出する手法の検討を行う．

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，東京都調布市調布ケ丘１－５－１，TEL:042-443-5664，Mail:tsuboi@uec.ac.jp

今回，我々は群知能分野における Ant Colony Optimization (ACO) アルゴリズムに着目している．ACO アルゴリズムは自然界におけるアリの群れとしての採餌行動をモデル化したものであり，さまざまな最適化問題に対して良い性能を示すメタヒューリスティクスであることが知られている．

既存のパターンマイニング技術でも度々課題となる項目として計算量の課題がある．網羅的にパターンを探索する場合，入力データが増えると指数関数的に計算量が増加する．ACO アルゴリズムから着想を得たアルゴリズムを構築することによって分散処理を行うなどの対応をすることで，有限な資源に沿った準最適な解を得ることに期待できる．また，分析対象が時系列データであるということは，時間経過に伴いパターンが変化することがあり得る．さらに，入力データにノイズが含まれるといった場合も想定される．このような問題に対しても，ACO アルゴリズムが有する優れた適応性・頑健性を発揮できるアルゴリズムの構築を試みる．

本稿の構成を次に示す．2 章では，パターンマイニングや ACO アルゴリズムに関する関連研究について述べる．3 章では，提案手法となる群知能メカニズムによる時系列階層構造を有するパターンの抽出手法について述べる．4 章では，提案手法に対してテストデータを用いた評価実験について述べる．5 章では．まとめと今後の課題について述べる．

2. 関連研究

大量のデータから頻出するパターンを抽出する手法がパターンマイニングである．データベース上から頻出するアイテムセットのパターンを抽出するためのアルゴリズムとして Apriori アルゴリズム [1] が有名である．Apriori アルゴリズムは小さいアイテムの集合から順に出現回数を数える幅優先探索である．あるアイテム集合が頻出とならない場合，その上位となるアイテム集合である，そのアイテムを含むより大きなアイテム集合について探索する必要がなくなる．このことより解の探索空間を減らしている．しかし，解の候補となるアイテム集合が生成されるたびに出現回数を数えるため，全ての入力データベースを調べなければならなくなる．そこで，入力データベースが膨大なものとなったときに，計算時間やメモリに問題が生じる．また，Apriori アルゴリズムでは，アイテムの時系列を

考慮していないといった問題が指摘された．

そこで，データベース上から時系列を考慮したアイテムのパターン抽出を行うために Apriori アルゴリズムを拡張した AprioriAll アルゴリズム [2] や，パターン抽出にあたって時間制約などの条件をつけることで計算処理の高速化を図った GSP アルゴリズム [3]，lattice（格子）という概念を用いて解の候補となるアイテムセットをグループ分割しそれぞれのグループをメモリ上に納めることで計算処理の高速化を図った SPADE アルゴリズム [4] が提案されている．これらのアルゴリズムは計算処理の高速化を図ったものであり，入力データを一様に等しく参照している．しかし，マーケティングにおける消費者の行動を考えると，流行や季節といったものに影響を受けて行動も変化するために，消費者の行動の変化に柔軟に対応できないという問題が生じる．

そこで着目したのが群知能分野の ACO アルゴリズムである．既存のデータマイニング手法では，膨大な計算コストを要したり実装が困難な状況において，実際に解が得られる手法として群知能の分野のアルゴリズムを応用する研究が行われている．実際に，群知能分野の ACO アルゴリズムをデータマイニングのクラスタリング手法に応用した AntMiner+アルゴリズム [5] は，分類結果がわかりやすい分類器でありながら精度が高く動的で分散した環境にも適したものである．

ACO アルゴリズムとは，自然界におけるアリの採餌行動をモデル化したもので，最適化問題である巡回セールスマン問題に対する解法として提案された．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，アリはフェロモンの濃度が濃い経路を好んだ経路選択をするというアリの行動を前提として考える．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．結果的に，アリが通過する頻度が少ない経路のフェロモンの濃度が薄くなる．残量するフェロモンの濃度が濃い経路が，最短経路問題の解となる．ACO アルゴリズムの特徴として，どのような環境に対しても解を生成することができる適応性や，環境の変化に対しても柔軟に解を再探索できる頑健性が挙げられる．

ACO アルゴリズムをパターンマイニングに応用した研究として Tamaki ら [6] の研究がある．Tamaki らは，センサが人の行動を読み取り反応することに着目し，連続した人の行動からセンサの隣接関係を，ACO アルゴリズムを応用することで推定する研究を行っている．人の行動が連続したものであり，それに対応してセンサが順次反応することからわかるように，時系列性が強い情報に対してマイニングを行っている．ACO アルゴリズムを応用したことで，複数の人が同時に行動した場合のノイズや，センサが故障や移動といった環境の変化にも対応できる．

このように，もともと最適化技術として提案された群知能分野のアルゴリズムであるが，単純な行動ルールに基づいた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，柔軟なシステムの構築が達成できる．本研究においても，ACO アルゴリズムが有する適応性や頑健性に着目し，群知能メカニズムによる時系列階層構造を有するパターンの抽出を試みる．

3. 群知能メカニズムによる時系列階層構造を有するパターンの抽出手法

提案アルゴリズムでは，アイテム集合 $I = \{i_1, i_2, \cdots, i_m\}$ からなる一つの時系列データ $D = \{d_1, d_2, \cdots, d_n\}$（ただし，$d \in I$）から，パターンとなる頻出するアイテムを順序関係を考慮しながら抽出する．

ACO アルゴリズムは主に，フェロモンを付加するフェーズとフェロモンを蒸発させるフェーズから成り立つ．本稿において提案するアルゴリズムにおいても ACO アルゴリズムと同様に，フェロモンを付加するフェーズ及びフェロモンを蒸発するフェーズからなる．また，現在の提案アルゴリズムは 2 層構造で構築しており，下位層では密に連結したパターンを抽出する一方で，上位層では下位層で抽出されたパターンを基に多少の粗さを許容しながら下位層より含まれる要素が多くなるようなパターンを抽出する．

3.1 仮想マップの用意

まず，フェロモンを付加するための環境として仮想マップ M_L, M_H を用意する．仮想マップ $M_L = (K_L, v_L(K_L))$ (ただし，$K_L \subset I$) および $M_H = (K_H, v_H(K_H))$(ただし，$K_H \subset I$) のそれぞれは，下位層，上位層それぞれのフェロモン量を記録した仮想マップである．仮想マップにおける K_L, K_H はパターンの候補を表し，$v_L(K_L), v_H(K_H)$ はそのパターンの候補に付与されたフェロモン量を表す．

3.2 下層におけるフェロモンの付加および蒸発

解候補を所持したアリエージェントが入力に対して探索を行い，探索結果をアリエージェントが評価し仮想マップに対して評価結果をフェロモンの付加を行う．

まず，入力である時系列データ D をタイムスパン T_L ごとに区切り，探索データセット $S_L(t)$ を作成する．各アリエージェントはそれぞれ好みとしてパターンの候補となる探索アイテム集合列 k_L が与えられ，探索データセット $S_L(t)$ 中に存在するパターン候補のアイテム集合列が連続して出現した個数 $C_{k_L}(t)$ を調査する．

各アリエージェントが探索するパターンの候補は，アイテム集合の中からパターン候補の大きさが 3 以下で構成される任意の集合 $\{\{i_1\}, \{i_2\}, \cdots, \{i_m\}, \{i_1, i_1\}, \cdots \{i_1, i_m\}, \cdots, \{i_m, i_m\}, \{i_1, i_1, i_1\}, \cdots \{i_m, i_m, i_m\}\}$ から式 1 の確率に従い設定する．

$$p(k'_L) = \frac{1}{\sum_{x=1}^{3} (\sum_I 1)^x} \tag{1}$$

これは，ACO アルゴリズムが巡回セールスマン問題に適用されるにおいて，アリエージェントがフェロモン量によらない探索を行いかつ，現在地点における周辺情報が等しいときにおける挙動と同様である．さらに選択された k'_L に対して，要素間にダミー要素 j_L を挿入した探索アイテム集合列 k_L に対して探索する．つまり，例えば，$k'_L = \{k_1, k_2, k_3\}$ であった場合，探索するアイテム集合列は $k_L = \{k_1, j_1, k_2, j_2, k_3\}$ となる．

一つの探索データセットごとに，l_L 匹のアリエージェントが探索を行い仮想マップの更新を式 4 に従い行う．

$$v(k) = v'(k)(1 - \rho_L) + C_k(t) \tag{2}$$

なお，ρ_L は蒸発率を表す．これは，蓄積されていたフェロモン $v'(k)$ に対して蒸発率 ρ_L を用いて減衰させた後に新たにアリエージェントが探索した結果をフェロモンという形で付加

している．フェロモンの蒸発率 ρ は情報更新の速さを表し，蒸発率が大きいほどフェロモンの蒸発は速くなり新しい情報に重みを置くようになる．一方，蒸発率が小さいほどフェロモンの蒸発は遅くなり古い情報を蓄積しやすくなる．

以上を探索データセットの続く限り行い，下層におけるパターンとなる連続して頻出するアイテム集合の抽出を行う．

3.3 上層におけるフェロモンの付加および蒸発

基本的な手順は下層と同様に，解候補を与えられたアリエージェントが入力に対して探索を行い，探索結果をアリエージェントが評価し仮想マップにフェロモンの付加を行う．

まず，入力である時系列データ D をタイムスパン T_H ごとに区切り，探索データセット $S_H(t)$ を作成する．各アリエージェントはそれぞれ好みとしてパターンの候補となる探索アイテム集合列 k_H が与えられ，探索データセット $S_H(t)$ の中に存在するパターンの候補のアイテム集合列が出現した回数 $C_{k_H}(t)$ を調査する．

上層における各アリエージェントが探索するパターンの解候補は，下層において抽出されたパターンの候補が記録された下層マップ $M_L = (K_L, v_L(K_L))$ を基に定められる．下層において抽出されたパターンの候補の集合である K_L の中から，任意に抽出される大きさが 3 以下の集合 $\{\{k_1\},\cdots,\{k_o\},\{k_1,k_1\},\cdots,\{k_o,k_o\},$ $\{k_1,k_1,k_1\},\cdots,\{k_o,k_o,k_o\}\}$ から式 3 の確率に従い選択する．

$$p(k'_H) = \frac{\prod_{k\in k2} v(k)}{\sum_{x=1}^{3}(\sum_K v_L(k_L))^x} \tag{3}$$

これは，ACO アルゴリズムが巡回セールスマン問題に適用されるにおいて，アリエージェントが過去に付加したフェロモン量のみに依存して探索する際の次の都市を選択する確率と同様である．さらに選択された k'_H に対して，要素間にダミー要素 j_H を挿入した探索アイテム集合列 k_H に対して探索する．つまり，例えば，$k'_H = \{k_1, k_2, k_3\}$ であった場合，探索するアイテム集合列は $k_H = \{k_1, j_1, k_2, j_2, k_3\}$ となる．

一つの探索データセットごとに，l_H 匹のアリエージェントが探索を行い仮想マップの更新を式 4 に従い行う．

$$v(k) = v'(k)(1-\rho_H) + C_k(t) \tag{4}$$

なお，ρ_H は上層における蒸発率を表す．これは下層と同様で，蓄積されていたフェロモン $v'(k)$ に対して蒸発率 ρ_H を用いて減衰させた後に新たにアリエージェントが探索した結果をフェロモンという形で付加している．フェロモンの蒸発率 ρ は情報更新の速さを表し，蒸発率が大きいほどフェロモンの蒸発は速くなり新しい情報に重みを置くようになる．一方，蒸発率が小さいほどフェロモンの蒸発は遅くなり古い情報を蓄積しやすくなる．

4. 評価実験

提案するアルゴリズムが想定したパターンを抽出できることを検証するために，テストデータを用いた評価実験を行う．

4.1 テストデータ

想定する解となるパターンを抽出できることを確認するために，抽出されるべきパターンを埋め込んだテストデータセットを生成する．登場するアイテム集合を $I = \{0, 1, 2, 3, \cdots, 99\}$ の 100 種類として，抽出されるべきパターンを $\{0, 1, 2, 3\}$ および $\{4, 5, 6, 7\}$ としたテストデータを生成する．

まず，抽出されるべきパターンに用いられるアイテム集合を除いた $\{8, 9, \cdots, 99\}$ のアイテム集合から一様乱数でテストデータ長分の乱数を時系列データとして生成する．今回の実験では，テストデータ長を 1000 とした．その後，時系列データに対して，抽出されるべきパターンである $\{0, 1, 2, 3\}$ および $\{4, 5, 6, 7\}$ を，それぞれ 20 個，10 個ずつ順番に埋め込む．パターンを埋め込む際に，すべてのパターンが等しくきれいに連続して出現することを防ぐために，あえてノイズを混入させるために，べき分布に従う乱数を生成し得られた乱数により埋め込むパターンとなるアイテムの間隔をあけてパターンの埋め込みを行う．このようにして得られるパターンが埋め込まれたデータを，入力となる時系列データとみなし実験を行う．

4.2 テストデータを用いた評価実験

生成したテストデータを用いて提案アルゴリズムの検証を行い，想定されるパターンを抽出できることを確かめる．今回の実験では，下層における探索のタイムスパンを $T_L = 5$ に，上層における探索のタイムスパンを $T_H = 10$ に設定した．また，本来それぞれの層におけるタイムスパンに従いフェロモンの蒸発を行うが，パターンを階層的に構築できることを確認するために，蒸発率は 0，つまりフェロモンの蒸発を実行しない環境で実験を行う．また，下層，上層それぞれにおいて，各タイムスパンごとに探索を行うアリエージェントの数は $l_L = l_h = 1000000$ として，実験を行った．

また，下層におけるダミー要素を $j_L = \{\emptyset\}$ として，上層におけるダミー要素を $j_H = \{i | i \in I \wedge |i| < 1\}$ とする．これは，下層においては探索する要素が連続で出現しているものについて探索を行う一方で，上層においては，探索アイテム集合の間に一つのみのノイズとなる他のアイテムが含まれる場合にも探索における回数に数えることを許す．

下層における仮想マップ M_L において付与されたフェロモン量が多い順に 10 個のパターンの候補および，想定した抽出されるべきパターンに関する候補を表 1 に示す．埋め込んだパターンの部分集合のうち，$\{0\}, \{2\}, \{3\}$ がパターンの候補に抽出されていることが確認できた．また，フェロモン量が 13 以上のものは要素が一つのもののみであったが，複数の要素を有するパターンの候補では $\{2, 3\}$ がフェロモン量 13，$\{0, 1\}$ がフェロモン量 11 と出現していた．

次に，上層における仮想マップ M_H において付与されたフェロモン量が多い順に 10 個のパターンの候補および，想定した抽出されるべきパターンに関する候補を表 2 に示す．埋め込んだパターンである $\{0, 1, 2, 3\}$ のさまざまな部分集合が抽出できているといえる．また，フェロモン量 6 には $\{\{4\}, \{5\}, \{6\}\}$ が出現していることも確認できた．

4.3 考察

実験より，下層においては想定したパターンの部分集合が取得できることがわかる．出現頻度の大小の通り $\{0, 1, 2, 3\}$ の部分集合が $\{4, 5, 6, 7\}$ の部分集合より多く出現している．抽出したいパターンの部分集合はフェロモン量が多い少ないの差はあれども，しっかりと出現していることが確認できた．さらに，上層における抽出パターンを見てみると，集合の集まりとして $\{0, 1, 2, 3\}$ の想定したパターンが抽出できたといえることが確認できる．一方 $\{4, 5, 6, 7\}$ の想定したパターンに対しては，まだ部分集合である $\{4, 5, 6\}$ と $\{5, 6, 7\}$ のように抽出されているが部分集合の集まりとしてすべてを含むパターンを抽出することができなかった．下層に対する上層をつくったことで，$\{0, 1, 2, 3\}$ が抽出できたように，$\{4, 5, 6, 7\}$ においても上層に対するさらなる上層を構築することで抽出ができる可能

表 1: 下層の仮想マップ上のフェロモン量が多い上位 10 候補および，想定したパターンに関する候補

候補	$v_L(k_L)$	候補	$v_L(k_L)$
{52}	23	{0}	20
{2}	20	{13}	20
{87}	20	{3}	19
{47}	19	{66}	18
{75}	17	{8}	16
{1}	12	{7}	12
{2,3}	12	{0,1}	11
{4}	10	{1,2}	9
{6}	9	{5}	8
{6,7}	7	{4,5}	6
{5,6}	6	{1,2,3}	5

表 2: 上層の仮想マップ上のフェロモン量が多い上位 10 候補および，想定したパターンに関する候補

候補	$v_H(k_H)$	候補	$v_H(k_H)$
{0}, {1}, {2}	31	{1}, {2}, {3}	26
{0}, {2}, {3}	13	{2}, {3}, {3}	11
{0}, {1}, {3}	9	{0}, {1, 2}, {3}	9
{0}, {75}, {0}	9	{0, 1}, {2}, {3}	8
{52}, {2}, {3}	7	{3}, {0, 1}, {3}	7
{4}, {5}, {6}	6	{0}, {1}	4
{2}, {3}	4	{2}, {3}, {4}	3
{5}, {6}, {7}	3	{5}, {6}	1
{3}, {5}, {6}	1	{4}, {5}	1
{3}, {4,5}, {6}	1		

性があると考える．

実験より，上層においては埋め込んだ数が多い $\{0, 1, 2, 3\}$ の方のパターンについて比較的きれいに抽出できているように考えられる．また，下層において抽出されるパターンに基づいて，上層の探索を行えているからこそ，上層における下層マップのフェロモン量の上位のものは下層のものと比べ，値が大きくなっている．集中した探索を行えていると考えられる，

5. おわりに

本稿では，群知能メカニズムの代表的なアルゴリズムである ACO アルゴリズムを基に，時系列データから頻出する連続するアイテムの組み合わせをパターンとして抽出する手法を提案している．さらに，アルゴリズムを下層と上層からなる階層構造を用いることで，それぞれの層に適合する粒度のパターンの抽出を試みている．提案するアルゴリズムに対して性能評価を行うために，抽出されるべきパターンを想定し，乱数上にパターンを埋め込んだテストデータを用いた実験を行った．結果として，出現頻度が多いパターンについては，比較的きれいにパターンの抽出が行えた．

今後の課題としては，出現頻度ごとにしっかりとパターンを抽出できる階層構造作りを行う．さらに，現在の提案モデルでは，ACO モデルの適応性の大きく影響するフェロモンの蒸発という要素が抜け落ちている．前半，後半で埋め込むパターンを変えた場合の時系列の変化にも対応するためにはフェロモンの蒸発という要素の導入も必要である．

また，実際のアルゴリズムの適用として，人の日常における

図 1: エージェントを動かす仮想空間の例

行動パターンの抽出を考える．まず，図 1 のような生活空間を仮想的に用意する．その仮想空間上を基本的な移動パターンを与えたエージェントを移動させ，時系列データの生成を行う．ここで生成された時系列データを入力データとして，提案アルゴリズムを用いて移動パターンの抽出を行う．他にも，アルゴリズムのマーケティングデータへの適用も考えられる．店舗における売れ方の時系列データをもとに，日におけるパターンであったり，週におけるパターンであったりとまちまちである．それらを自動的に抽出できるように工夫していくことも考えられる．

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] Martens David, et al . "Classification with ant colony optimization" IEEE Transactions on Evolutionary Computation 11.5, pp.651-665,2007.

[6] 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.