Vol.2010-MPS-78 No.11
2010/5/21

局所分枝限定探索による概念プール更新操作に基づく萌芽的概念のボトムアップ抽出

中島 健志[†1] 原口 誠[†1] 大久保 好章[†1]

萌芽的概念の抽出に向けて，本稿では，簡潔なレア概念の Top-N 抽出問題のための，分枝限定深さ優先探索による概念プールの効率的な更新処理を核とする高速なボトムアップアルゴリズムを与える．部品概念群を格納した概念プールに対して，探索の深さを限定した上で，部品概念の組み合わせを深さ優先探索し，評価値が上位 N のレア概念を求め，これを新たな概念プールとする．こうした概念プールに対して同様の局所的な Top-N 探索を繰り返すことで一種のビーム探索を実現し，準最適な Top-N レア概念の高速抽出を試みる．

Bottom-Up Extraction of Concise Rare Concepts by Excavation of Concept Pools with Local Branch-Bound Searches

Takeshi NAKAJIMA , Makoto HARAGUCHI and Yoshiaki OKUBO

We present in this paper an algorithm for finding Top-N *Concise Rare Concepts* (CRCs) with local branch-and-bound searches. A CRC is a concept whose extent is smaller (not frequent) and intent consists of a small number of general attributes, and is therefore opposite to colossal patterns. In order to efficiently extract those concepts especially from a large scale dataset, we design a bottom-up algorithm with branch-and-bound prunings. In the algorithm, we iterate a depth-bounded *local Top-N rare concept search* with a *concept pool*. At each iteration stage, the concepts in the pool are combined in depth-first manner so that we can obtain larger Top-N rare concepts. The concept pool is updated by replacing the original pool with the newly obtained Top-N concepts, then the same procedure is iterated until no update on pools is observed. In this sence, a kind of *beam-search* is carried out in our algorithm.

1. はじめに

データマイニング研究の主要なひとつのテーマとして，**飽和アイテム集合**[2),3)]，あるいは，それと等価な**形式概念**[1)] の抽出・列挙問題が注目されて久しい．これらの研究では主に，生起頻度が比較的大きな**頻出パターン**を抽出のターゲットとしてきた．頻出パターンの中でも，アイテム数が少数な場合はその検出タスクは比較的に容易であり，頻出だが**長大なパターン** (Colossal Pattern) の発見問題も考察されている[4)]．本稿ではこれらと対象的に，

『稀だが少数の一般的なアイテムから構成されるパターンは意外性に富む』

との考えに基づき，非頻出でかつ内包において一般的な概念パターン，すなわち，**萌芽的概念抽出**を試みる．

著者等はこれまでの研究により，上述した意味での萌芽的概念を**簡潔なレア概念** (Concise Rare Concept) として定式化した[5),6)]．そこでは，概念の形成過程に注目しながら，高い相関を示す冗長な属性を含む内包 (あるいはその生成元) を制約により排除し，その制約のもとで，内包の一般性が上位 N であるレア概念を抽出ターゲットとする．一般に，レア概念は大きな内包を有することから，その意味解釈が容易でないが，概念形成過程の制約により冗長な内包が排除され，結果として，比較的小さな内包を有するレア概念が抽出される．抽出アルゴリズムは，形式概念束において最上位に位置する概念，すなわち，すべての個体が属する最も一般的な概念を起点とし，その下位概念を深さ優先で探索するトップダウン戦略を用いたものであり，制約および目的関数の単調性に基づく枝刈りを利用して，ターゲットを効率良く抽出する．例えば，11,000 の Web 文書 (索引語数およそ 1,200) を用いた計算機実験により，数秒のオーダで簡潔なレア概念の抽出が可能であることを確認している．

一方で，概念形成過程の制約に関するパラメータ設定によっては，解が存在しない場合もあり，その調整はさほど容易とは言えない．また，様々な萌芽的概念を抽出するためには，より大規模なデータへの適用を見据えた上で，さらに高速なアルゴリズムの開発が強く望まれる．そこで本稿では，これらの問題に対処すべく，新たなレア概念抽出手法について考察する．

具体的に述べると，これまで制約により間接的に制御していたレア概念の内包サイズを，目的関数内で陽に考慮することで，直接最適化の対象とする．これにより，解を得るための

†1 北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

パラメータ設定の困難さが解消される．また，レア概念は概念束中の下方に位置することから，特に大規模データでは個体集合 (外延) の拡張処理を基本とするボトムアップアプローチがより適しているとの期待が持てる．ここでは，部品となる概念群を格納した**概念プール**を考え，プール中の部品概念を結合することでより大きな概念を生成する処理を繰り返す．ある概念プールに対して，探索の深さを限定した上で，部品概念の組み合わせを探索し，評価値が上位 N のレア概念を求め，これを新たな概念プールとする．こうした概念プールに対して同様の局所的な Top-N 探索を繰り返すことで，一種の**ビーム探索** (Beam-Search) を実現する．

2. 準　　備

個体の集合 G，および，**属性**の集合 M に対して，関係 $I \subseteq G \times M$ を考える．この時，タプル (G, M, I) を**形式文脈** (*Formal Context*) と呼ぶ．$(g, m) \in I$ の時，個体 g は属性 m を有すると言う．

形式文脈 (G, M, I) において，写像 $\varphi : 2^G \to 2^M$ および $\psi : 2^M \to 2^G$ を考える．個体集合 $X \subseteq \mathcal{G}$ と属性集合 $Y \subseteq \mathcal{M}$ について，

$$\varphi(X) = \{m \in M \mid \forall x \in X, \quad (x, m) \in I\},$$
$$\psi(Y) = \{g \in G \mid \forall y \in Y, \quad (g, y) \in I\}$$

とする．

これら写像のもと，個体集合 $X \subseteq G$ と属性集合 $Y \subseteq M$ について，$\varphi(X) = Y$ かつ $\psi(Y) = X$ が成り立つ時，X と Y の組 $C = (X, Y)$ を**形式概念** (*Formal Concept*) [1] と定める．ここで，X と Y をそれぞれ C の**外延** (*Extent*)，および，**内包** (*Intent*) と呼ぶ．なお，以下の議論では，特に断りのない限り，単に概念と言った場合は，形式概念を指すものとする．

概念 $C = (X, Y)$ および $C' = (X', Y')$ について，$X \subseteq X'$ $(Y \supseteq Y')$ である時，かつ，その時に限り C と C' 間に順序関係を定め，これを $C \preceq C'$ と表記する．この時，C は C' の**特殊概念**, 逆に，C' は C の**汎化概念**と呼ぶ．所与の形式文脈におけるすべての形式概念の集合を $\mathcal{FC}$ とすると，順序関係 $\preceq$ のもと，$(\mathcal{FC}, \preceq)$ は束を構成し，これを**形式概念束** (*Formal Concept Lattice*) と呼ぶ．

形式文脈 (G, M, I) における個体と属性間の二項関係 I は，パターンマイニングにおける，トランザクションデータに他ならない．つまり，個体をトランザクション，属性をアイテムと考えれば，形式概念 (X, Y) の内包 Y は**飽和アイテム集合** (*Closed Itemset*)[3] に，また，外延 X は (飽和) アイテム集合 Y を含むトランザクション集合に対応する．よって，外延 X の大きさ $|X|$ は，アイテム集合 Y の**頻度**に，また，$|X|/|G|$ は Y の**サポート**に一致する．以下では，Y のサポートを $support(Y)$ で参照する．

3. 簡潔なレア概念

先行研究にならい，本稿においても，萌芽的概念として抽出すべきレア概念に対して

『その内包は少数の一般的な属性から成る』

ことを要請し，これを**簡潔なレア概念**と呼ぶ．

より正確な議論を進めるために，まず，**レア概念**を次の通り定義する．

定義 3.1　レア概念 (Rare Concept)
$C = (X, Y)$ を形式概念，R を**レアネス閾値**とする．

$$|X|/|G| = support(Y) \leq R$$

である時，C を R-**レア概念**と呼ぶ． ■

R-レア概念とは，概念束中で閾値 R が定める境界の下方に位置する概念を指している．

次に，概念の**簡潔性**について議論する．概念の意味解釈はその内包により与えられることから，簡潔性は内包に基づいて定義するのが自然である．ここでは，内包の**サイズ**と**一般性**のふたつの点から簡潔性を考える．

内包サイズが小さい程，簡潔な意味解釈が期待できることは言うまでもない．よってその度合を測る**コンパクトネス**を次の通り定義する．

定義 3.2　内包のコンパクトネス (Compactness)
$C = (X, Y)$ を形式文脈 (G, M, I) における概念とする．この時，C の内包 Y のコンパクトネスを $compactness(Y)$ と表し，

$$compactness(Y) = \frac{|M| - |Y|}{|M|}$$

とする． ■

内包の**一般性**は，それを構成する属性のサポート (頻度) に基づいて測るものとする．ここでは，内包中の属性の最小サポートにより一般性を定める．

定義 3.3　内包の一般性 (Generality)
$C = (X, Y)$ を形式文脈 (G, M, I) における概念とする．この時，C の内包 Y の一般性を $generality(Y)$ と表し，

$$generality(Y) = \min_{y \in Y} \{support(\{y\})\}$$

とする． ■

これら尺度に基づいて，内包の**簡潔性**を次の通り定める．

定義 3.4　内包の簡潔性 (Conciseness)
$C = (X, Y)$ を形式文脈 (G, M, I) における概念とする．この時，内包 Y の簡潔性を $conciseness(Y)$ と表し，

$$conciseness(Y) = \alpha \cdot compactness(Y) + \beta \cdot generality(Y)$$

とする．ここで，α と β はそれぞれ，コンパクトネスおよび一般性を考慮する度合を制御する非負の重みパラメータである． ■

以下では，簡潔な概念とは，簡潔な内包を有する概念を指すものとする．

形式概念の定義より，$(X, Y) \preceq (X', Y')$ である概念間には，$Y \supseteq Y'$ なる関係がある．よって，内包の一般性の定義から，$generality(Y) \leq generality(Y')$ が成り立つ．コンパクトネスについては，$compactness(Y) \leq compactness(Y')$ が明らかであるから，簡潔性についても，$conciseness(Y) \leq conciseness(Y')$ となり，すなわち，内包の縮小に伴い簡潔性は単調に増加することがわかる．

以上の議論を踏まえ，本稿で扱う簡潔なレア概念抽出問題を次の通り定義する．

定義 3.5　簡潔なレア概念の Top-N 抽出問題
(G, M, I) を形式文脈，R をレアネス閾値とする．この時，以下の条件を満たす形式概念 $C = (X, Y)$ を抽出する問題を，**簡潔な R-レア概念の *Top-N* 抽出問題**と呼ぶ．

制約 (レアネス)： C は R-レア概念である．

目的関数 (簡潔性)： C の内包 Y の簡潔性は，任意の R-レア概念中で上位 N 以内である． ■

次章では，簡潔なレア概念の抽出アルゴリズムについて議論する．

4. 概念プールの更新に基づく Top-N レア概念抽出

4.1 探索の基本戦略

レア概念は形式概念束の下方に位置するため，特に大規模データを扱う場合は，概念束の下方から上方へ探索を進める**ボトムアップアプローチ**が有望に思える．よって，ここでは，個体集合の拡張処理を基本とするアルゴリズムを与える．

概念束中に存在する概念の総数は膨大であることがよく知られているが，束中でのそれらの分布には大きな偏りが見られ，特に下方に位置する概念は極めて多い．上述した Top-N レア概念抽出問題の最適解を見つけるには，この様な膨大な概念が密集する領域を探索する必要があるが，大規模データに対してそれをまともに行なうのは現実的ではない．そこで本稿では，ある種の**ビーム探索** (*Beam-Search*) により，大規模データにおいても現実的な計算時間で準最適解を求めるアルゴリズムを提案する．

具体的には，部品となる概念群を保持する**概念プール**を用意し，探索の対象を，それら部品の結合によって得られる概念に限定する．ある概念プールをもとに，探索木の深さを限定することで，**局所的な Top-N レア概念探索**を行ない，その抽出結果を新たな概念プールとする．この様な局所的 Top-N 探索による概念プールの更新処理を，初期概念プールから順次繰り返し，更新が観測されなくなった時点の概念プールを準最適解として出力する．

まとめると，探索の基本戦略は次の通りとなる．

入力：N, R, D, α, β.
出力：準最適な簡潔性上位 N の R-レア概念集合.
1. 概念プールの初期化：
　$P \leftarrow$ 簡潔性が上位 N のプリミティブ概念集合
2. 局所的 **Top-N** レア概念探索：

$L \leftarrow \mathrm{LocalTopNSearch}(P)$

3. 停止条件判定：

$L = P$ ならば，L を出力して停止．

$L \neq P$ ならば，$P \leftarrow L$ として **2** へ戻る．

ここで，LocalTopNSearch は，概念プール P を入力とし，P 中の概念を高々 D 結合することで構成可能な Top-N レア概念を出力する手続きである．

以下では，主要な処理の詳細について議論する．

4.2 初期概念プール

概念プール中の概念群は，局所的探索における結合操作の対象となる．よって，初期概念プールには，**プリミティブな概念**の内，レアでかつ簡潔性が上位 N であるものを格納する．ここで，プリミティブ概念とは，単一の個体から求まる概念であり，正確には次の通り定義される．

定義 4.1 プリミティブ概念 (Primitive Concept)

(G, M, I) を形式文脈とする．個体 $x \in G$ について，

$$(\psi(\phi(\{x\})), \phi(\{x\}))$$

なる概念を，x から構成される**プリミティブ概念**と呼び，これを $prim(x)$ と表す． ■

いま，G 中の個体から構成される R-レアなプリミティブ概念の集合を $RarePrim(G)$ とすると，初期概念プールは，

$$\{C \mid C \in RarePrim(G) \wedge C \text{ の簡潔性は } RarePrim(G) \text{ 中で上位 } N\}$$

となる．

4.3 局所的 Top-N レア概念探索による概念プールの更新

所与の概念プール P の更新は，P 中の部品概念を結合することで得られるレア概念の内，簡潔性が上位 N であるものを求めることで行なう．結合操作は，後に述べる枝刈り規則を有効に利用するために，深さ優先で行なうこととする．

プール中の部品概念の結合処理は，**集合列挙木**に基づいて行なうことができる．いま，集合 $S = \{s_1, \ldots, s_{|S|}\}$ 上に，ある全順序 $s_1 \prec \cdots \prec s_{|S|}$ を仮定し，S の部分集合 A 中の要素はこの順序に従ってソーティングされているものとする．また，S の部分集合 $A_i = \{s_{i_1}, \ldots, s_{i_k}\}$ について，A_i の最終要素 s_{i_k} を $tail(A_i)$ で参照する．さらに，先頭の ℓ 要素の集合 $\{s_{i_1}, \ldots, s_{i_\ell}\}$ を，A_i の ℓ-**接頭辞**と呼び，$prefix(A_i, \ell)$ で参照する．特に，$prefix(A_i, 0) = \phi$ とする．

ここで，次の通り定義される 2^S 上の半順序 $\prec_S$ を導入する．

定義 4.2 2^S 上の半順序

S の部分集合 A_i および A_j を考える．A_i が A_j の $|A_i|$-接頭辞である，すなわち，$A_i = prefix(A_j, |A_i|)$ である時，かつ，その時に限り A_i は A_j の**前者**であると言い，これを $A_i \prec_S A_j$ と表わす．また，A_i が A_j の直接の前者である時，A_j を A_i の**子供**と呼ぶ． ■

半順序集合 $(2^S, \prec_S)$ は，空集合 ϕ をルートノードとする木を構成し，これを**集合列挙木** (*Set Enumeration Tree*) と呼ぶ．集合列挙木中の各内部ノードに対応する $A \subseteq S$ について，その子供を得るには，単に，$tail(A) \prec s$ なる任意の要素 s を用いて $A \cup \{s\}$ とすればよい．この様に，空集合を起点にこうした処理を繰り返すことで，S のすべての部分集合を，**機械的**，かつ，**重複無く**列挙することができる．

これより，概念プール P 中の部品概念を，P の集合列挙木に従って結合することで，すべての可能な結合処理を行えることがわかる．先に触れた通り，ここでは特に深さ優先で集合列挙木を探索するものとする．

その際，探索の深さ，すなわち，結合する部品数に上限 D を設けることで，探索すべき範囲が不用に広がることを避ける．概念プールと探索の深さ上限により，探索範囲が概念束中のある一部分に限定されることから，ここでの探索は，**局所的な** *Top-N* **探索**に相当している．

形式概念の性質から，概念の結合処理の結果得られる新たな概念は，次の通りとなる．いま，$C_i = (X_i, Y_i)$ および $C_j = (X_j, Y_j)$ を結合して得られる概念を C_{ij} とすると，C_{ij} は次式で与えられる．

$$\begin{aligned} C_{ij} &= (\psi(\phi(X_i \cup X_j)), Y_i \cap Y_j) \\ &= (\psi(Y_i \cap Y_j), Y_i \cap Y_j). \end{aligned}$$

プール中の概念の結合処理に伴い，概念のレアネスは単調に減少することから，R-レアでない概念が得られた時点で，それ以降の結合処理を打ち切って，バックトラックすることができる．

また，局所的 Top-N 探索においては，暫定解の最小評価値 (簡潔性) に基づいて不要な探索枝を刈ることも可能である．先に述べた通り，内包が縮小するに伴い簡潔性は単調増加する．形式概念の性質より，このことは，外延の拡張に伴い簡潔性が単調増加することを示している．いま，これまでの探索で暫定的に見つかった Top-N 概念の簡潔性最小値を δ とする．概念の結合操作により簡潔性は単調に増加するから，探索のある時点で得られた概念 C に対して，その先の探索により得られる概念の簡潔性上限値 σ を見積もることができれば，δ と σ を比較し，$\sigma < \delta$ である場合は，概念 C 以降の探索で Top-N になり得る概念は得られないことがわかる．よって C を起点とする結合処理を安全に枝刈りすることができる．

こうした上限値 σ は，次の通り見積もることができる．いま，概念 $C = (X, Y)$ にさらに部品概念を結合することを考える．結合処理によって得られる概念を $C' = (X', Y')$ とすると，$Y \supseteq Y'$ であるから，探索木の葉に相当する概念の内包は，属性 $y \in Y$ について，

$$Y' = \phi(\psi(\{y\}))$$

で与えられる．よって，これら概念の内包の内，最も高い簡潔性の値を σ にとれば，それが，C を起点とする結合操作によって得られる概念の上限評価値を与えることになる．すなわち，

$$\sigma = \max_{y \in Y} \{conciseness(\phi(\psi(\{y\})))\}$$

とすればよい．

上述した枝刈り機構を組み込んだ局所的 Top-N レア概念探索アルゴリズムの擬似コードを図 1 に示す．ここで，Combine(C_i, C_j) は，概念 C_i と C_j の結合概念を返す手続き，Intent(C) は，概念 C の内包を返す手続き，LowestValue(L) は，L 中の概念の簡潔性最小値を返す手続きとする．

5. おわりに

萌芽的概念の抽出を目的として，本稿では，少数の一般的な属性から成る内包を有するレ

Procedure LocalTopNSearch(P) :
// 概念プール P 中の部品概念を，高々 D 結合することで得られる
// Top-N R-レア概念を返す.
  $L \leftarrow P$ // 暫定 Top-N リスト L の初期化
  for each $i \in \{1, \ldots, |P|\}$ **do**
    begin
      TopNSearchSub(L, P, $P[i]$, $i+1$, $D-1$);
    end
  return (L);

Procedure TopNSearchSub(L, P, C, k, d) :
// 概念プール P 中の k 番目以降の部品概念を，概念 C に高々 d 結合する
// ことで得られる Top-N R-レア概念を L に格納する.
  if $d = 0$ **then** return;
  for each $i \in \{k, \ldots, |P|\}$ **do**
    begin
      $NewC \rightarrow$ Combine(C, $P[i]$) ;
      $\sigma = \max_{y \in \mathrm{Intent}(NewC)} \{conciseness(\phi(\psi(\{y\})))\}$;
      if $NewC$ is R-rare and $\sigma \geq$ NthValue(L) **then**
        begin
          TopNListUpdate(L, $NewC$);
          TopNSearchSub(L, P, $NewC$, $i+1$, $d-1$);
        end
      endif
    end

Procedure TopNListUpdate(L, C) :
// 概念 C の追加に伴い，暫定 Top-N リスト L を更新する.
  $L \leftarrow L \cup \{C\}$;
  Remove all concepts from L with M-th conciseness-value such that $N < M$;

図 1 局所的 Top-N レア概念探索アルゴリズム

ア概念を抽出するボトムアップアルゴリズムについて考察した．特に大規模データに対応すべく，ここでは，局所的な Top-N レア概念探索による概念プールの更新処理を繰り返すことで，ビーム探索により簡潔性の評価が Top-N であるレア概念を抽出する．

これまで，局所的探索による概念プール更新処理を基本としたビーム探索アプローチの有効性を，数千オーダのデータに対する予備実験において確認している．本アプローチの有効性は，大規模なデータに対してより顕著に現れると予想されることから，現在，個体数が数十万オーダの大規模データに対する計算機実験を計画・準備中である．その結果については別途報告をしたい．

また，抽出される概念の品質を萌芽性の観点からさらに分析し，それを制約および目的関数に反映させることも重要な課題である．

参 考 文 献

1) B. Ganter and R. Wille, Formal Concept Analysis - Mathematical Foundations, Springer, 284 pages, 1999.
2) T. Uno, M. Kiyomi and H. Arimura. LCM ver. 2: Efficient Mining Algorithm for Frequent/Closed/Maximal Itemsets. Proc. of IEEE ICDM'04 Workshop - FIMI'04, http://sunsite.informatik.rwth-aachen.de/verb+Publications/CEUR-WS//Vol-126/, 2004.
3) N. Pasquier, Y. Bastide, R. Taouil and L. Lakhal, Efficient Mining of Association Rules Using Closed Itemset Lattices, Information Systems, 24(1), pp. 25 - 46, 1999.
4) F. Zhu, X. Yan, J. Han, P. S. Yu and H. Cheng, Mining Colossal Frequent Patterns by Core Pattern Fusion, Proc. of IEEE 23rd International Conference on Data Engineering - ICDE'07, pp. 1 - 10, 2007.
5) 中島 健志・原口誠・大久保好章，萌芽的閉包を枚挙する分枝限定法について，情報処理学会研究報告，Vol. 2009-MPS-76 No. 14 (Vol. 2009-BIO-19 No. 14), 2009.
6) Y. Okubo and M. Haraguchi, An Algorithm for Extracting Rare Concepts with Concise Intents, Proceedings of the 8th International Conference on Formal Concept Analysis - ICFCA'10, Springer-LNAI 5986, pp. 145 - 160, 2010.