1A2-1

萌芽的概念抽出のための 局所分枝限定探索を用いた概念プール掘削法

Finding Concise Rare Concepts with Excavation of Pattern Pools Based on Local Branch-and-Bound Searches

中島 健志 Takeshi NAKAJIMA　原口 誠 Makoto HARAGUCHI　大久保 好章 Yoshiaki OKUBO

北海道大学大学院情報科学研究科コンピュータサイエンス専攻
Division of Computer Science, Garduate School of Information Science and Technology, Hokkaido University

In this paper, we present 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.

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い頻出パターンが抽出のターゲットとされてきた．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]．本研究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．

その具体例として，特許文書に記述された『カーテン型スピーカ』が挙げられる．これは,『カーテン』に関する属性，および,『スピーカ』に関する属性を持つが，これら属性を同時に持ち合わせる概念は非常に稀であり，大変意外性に富んだものである．この様な稀有な属性の組合せを持つ概念 (発明) は，今後のさらなる発展の可能性や，新たな発明を呼ぶ一助となる可能性を示唆するという意味で，萌芽的な重要なものと考えられるであろう．

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，その意味解釈は容易ではないが，概念形成過程の制約によって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，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順次調べ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る．また，様々な萌芽的概念を抽出するためには，さらに大規模なデータへの対応も不可欠である．そこで本研究では，これらの問題に対処すべく，新たな簡潔なレア概念抽出法について考察する．

より具体的に述べると，これまで概念形成過程の制約により制御していたレア概念の内包サイズを，直接制約として与えて単純化する．これにより，解を得るための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期待できる．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**を実現する．

2. 準備

個体 (*object*) の集合 G，および，**属性** (*attribute*) 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．個体 g が有する属性の集合 $\{m \in \mathcal{M} \mid (g, m) \in I\}$ を，$M(g)$ で参照する．

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

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

連絡先: 原口 誠
北海道大学大学院情報科学研究科
〒 060-0814 札幌市北区北 14 条西 9 丁目
TEL : 011-706-7106(ダイヤルイン)
E-mail : mh@ist.hokudai.ac.jp

とする．特に，属性集合 Y の**サポート**を $sup(Y) = |\psi(Y)|/|G|$ と定める．

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 $FC = (X, Y)$ を**形式概念** (*Formal Concept*) [1] と定める．ここで，X と Y をそれぞれ FC の**外延** (*extent*)，および，**内包** (*intent*) と呼ぶ．

形式概念 $FC = (X, Y)$ および $FC' = (X', Y')$ について，$X \subseteq X'$ ($Y \supseteq Y'$) である時，かつ，その時に限り FC と FC' 間に順序関係を定め，これを $FC \preceq FC'$ と表記する．この時，FC は FC' の**特殊概念**, 逆に，FC' は FC 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．

3. 簡潔なレア概念の Top-N 抽出問題

本研究では，萌芽的概念として抽出すべきレア概念に対し

『その内包は一般的かつ少数の属性から成るべき』

なる特徴を有することを要請する．

形式文脈 (G, M, I) における概念の**レアネス**は，その外延サイズにより規定する．すなわち，R をレアネス閾値とすると，形式概念 $C = (X, Y)$ が $|X| \leq R$ を満たす時，C を R-**レアな概念**と呼ぶ．

内包の**一般性**は，それを構成する属性の頻度に基づいて測るものとする．ここでは，内包中の属性のサポート値の平均により一般性を定める．すなわち，形式概念 $C = (X, Y)$ について，その内包 Y の一般性を $generality(Y)$ と表し，

$$generality(Y) = \frac{1}{|Y|} \sum_{y \in Y} sup(\{y\})$$

と定義する．

これら尺度に基づいて，本研究におけるレア概念抽出問題を次の通り定める．

定義 3..1 (簡潔なレア形式概念の Top-N 抽出問題)
(G, M, I) を形式文脈，R をレアネス閾値，L を内包最大サイズ閾値とする．この時，(G, M, I) のもとで，以下の条件を満足する形式概念 (X, Y) を抽出する問題を，**簡潔な** R-**レア形式概念の** *Top-N* **枚挙問題**と呼ぶ．

レアネスおよびサイズ (制約)： (X, Y) は R-レアであり，かつ，その内包サイズは L 以下である．

一般性 (目的関数)： 内包 Y の一般性は，任意の R-レア概念中上位 N 以内である． ■

4. Top-N レア概念抽出アルゴリズム概要

レア概念は形式概念束の下方に位置するため，大規模データを扱う場合は特に，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**が有望に思えることから，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．

概念束中には膨大数の概念が存在す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 枚挙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．そこで本研究では，部品となる概念を保持する**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．

より具体的には，初期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 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．ここで，各個体 $x \in G$ について，$(\psi(\phi(\{x\})), \phi(\{x\}))$ を**プリミティブ概念**と呼び，初期概念プールは，R-レアで，かつ，一般性が上位 N であるプリミティブ概念からなるものとする．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，プールの更新が観測されなくなるまで繰り返すことで，一種の**ビーム探索**を実現する．

以上をまとめると，Top-N レア概念の抽出処理の主要な流れは次の通りとなる．

入力： R, L, N, D.
出力：一般性が上位 N の R-レア概念集合．

1. 概念プールの初期化：
$P \leftarrow \{$ プリミティブ概念 C |
C は R-レア $\wedge$ 一般性が上位 $N\}$

2. 局所的 Top-N 探索：
$L \leftarrow$ `LocalTopN_Search`(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，`LocalTopN_Search` は，概念プール P を入力とし，探索の深さ上限 D のもとで P 中の概念を結合することで構成可能な Top-N レア概念を出力する関数である．

局所的 Top-N 探索においては，暫定解の最小評価値 γ に基づいて不要な探索枝を刈ることが可能である．定義より，内包が小さくなるにつれて，一般性は単調増加する．形式概念の性質より，$(X, Y) \preceq (X', Y')$ である概念間には，$Y \supseteq Y'$ なる関係があるので，(X, Y) を拡張して得られる概念の一般性は，高々

$$max_{y \in Y}\{generality(\phi(\psi(\{y\})))\}$$

となる．よって，この値が γ に満たない場合は，(X, Y) の拡張を行なっても Top-N となる概念は得られないことがわかるため，そうした探索は安全に枝刈ることができる．

5. 実験

上述したアルゴリズムを `Java` で実装し，Intel Xeon E5520(2.27GHz)・主記憶 1 GB の PC 上で実験を行なった．データには，NTCIR-4 にて提供された特許文書群を用い，名詞のみを属性 (索引語) として用いた．文書総数は 4,000 である．

レアネス閾値 $R = 50$，内包サイズ閾値 $L = 3$，局所探索深さ上限 $D = 4$ のもとで，Top-30 のレア概念抽出を行なった結果，100 秒程度の計算時間で結果を得ることができた．得られたレア概念中には，著者らの主観に照らして萌芽的と言えそうなものも含まれているが，この点については，さらなる考察が必要である．

6. おわりに

萌芽的概念の抽出に向けて，本研究では，少数の一般的な属性から成る内包を有するレア概念の抽出問題について議論し，

概念の結合処理を基本操作とするボトムアップ抽出アルゴリズムを設計した．特に，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，ビーム探索により一般性の評価が Top-N であるレア概念を抽出する点が特徴である．

今後の重要な課題として，より大規模なデータへの適用を見据えたアルゴリズムのさらなる効率化，および，レア概念の評価指標の改良などが挙げられる．

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.