2008－MPS－71 (4)
2008／9／18

内包の核を考慮した疑似形式概念の Top-N 抽出

大久保 好章　　原口 誠
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

概要：本稿では，形式概念に基づくクラスタ抽出において生じる，外延の重複問題を緩和すべく，疑似形式概念について考察する．核となる主要な属性集合と，それ以外の副次的な属性集合から構成される内包を用いて定義される疑似形式概念を提案し，その Top-N 抽出アルゴリズムについて議論する．

Extracting Top-N Pseudo-Formal Concepts with Core Intents

Yoshiaki OKUBO and Makoto HARAGUCHI
Division of Computer Science, Graduate School of Information Science and Technology
Hokkaido University

Abstract: We propose in this paper a notion of *Pseudo-Formal Concept*(*PseudoFC* in short). A *PseudoFC* can be viewed as a natural approximation of formal concepts. It covers several formal concepts as its *majorities* and can work as a representative of them. Such a *PseudoFC* is defined as a tuple $(O, F \cup S)$, where O is a closed set of objects, F a set of *primary features*, S a set of *secondary features*. Then, the concept tells us that 1) all of the objects in O are associated with the primary features F and 2) for each secondary feature $f \in S$, a majority of O is also associated with f. Therefore, O can be characterized not only *exactly* by F but also *flexibly* by $F \cup \{f\}$ for each secondary feature f. Our task is formalized as a problem of finding *Top-N τ-Pseudo Formal Concepts*. The targets can be extracted based on *a depth-first branch-and-bound clique search* with some pruning rules.

1 はじめに

著者等はこれまで，文献 [原口 02] をその始まりとして，クリーク探索に基づくクラスタ抽出の研究を行なってきた．近年は，クラスタの**解釈・意味付け**が重要であるとの立場のもと，**形式概念解析** (*Formal Concept Analysis*) [1] の枠組で議論している [3, 4]．

文献 [3, 4] では，クラスタ抽出問題を，内包に関する制約を満たし，かつ，外延の評価値が Top-N の形式概念を求める問題として定式化し，その計算アルゴリズムを設計・実装した．それは著者らの従来アルゴリズムの拡張であり，分岐限定を利用した効率良い探索が可能である．

こうした形式概念に基づくクラスタ抽出において，重複の大きなクラスタが Top-N の大部分を占め，得られたクラスタ間に大きな違いを見出せない場合がしばしば観測される．同様な問題は文献 [5] でも議論され，そこでは**疑似クリーク** (*Pseudo-Clique*) の考えを導入することで，その緩和を試みた．本稿では，形式概念に基づくクラスタ抽出における重複問題を緩和するために，**疑似形式概念** (*Pseudo-Formal Concept*) を提案し，その計算アルゴリズムについて議論する．疑似形式概念は，外延の違いが許容誤差以内の形式概念群の代表元と捉えることができる．特に，疑似形式概念の内包は，すべての個体が共有する**主要属性**の集合 (核と呼ぶ) と，大部分の個体が有する**副次属性**の集合から構成される．こうした属性の区別によって，より自然な外延の特徴付けが可能となる．

連絡先　〒 060-0814　札幌市北区北 14 条西 9 丁目
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
TEL: 011-706-7161 (FAX 兼用)
E-mail: {yoshiaki,mh}@ist.hokudai.ac.jp

2 準備

V を**節点集合**，$E \subseteq V \times V$ を**枝集合**とする無向グラフ G を $G = (V, E)$ と表す．グラフ G において，節点 $v \in V$ と隣接する節点の集合を $N_G(v)$ で表し，その要素 (節点) 数，すなわち，$|N_G(v)|$ を G における v の**次数**と言う．これを $degree_G(v)$ で参照する場合もある．なお，文脈上明らかな場合は，単に $N(v)$ や $degree(v)$ と略記する．

グラフ G の任意の (異なる) 節点間に辺が存在する時，G を**完全グラフ**と呼ぶ．

グラフ $G = (V, E)$ おいて，V の部分集合を V' とする．$G' = (V', E \cap (V' \times V'))$ で定義されるグラフを，G の**部分グラフ**と呼び，$G(V')$ と表記する．特に，G' が完全グラフである時，それは**クリーク**と呼ばれ，単に，その構成節点集合 V' で表すものとする．また，その**サイズ**を $|V'|$ で定める．G のクリーク Q と Q' が，$Q \subseteq Q'$ の関係にある時，Q' を Q の**拡張** (*extension*) と呼ぶ．G のクリークのうち，包含関係のもとで極大なものを，**極大クリーク**と呼ぶ．特に，サイズが最大である極大クリークは**最大クリーク**と呼ばれる．一般に，最大クリークは一意に決まらないことに注意する．

クリーク Q について，Q の任意の頂点と隣接する頂点を，Q の**拡張可能頂点**と呼び，こうした頂点集合を $cand(Q)$ で参照する．任意の $v \in cand(Q)$ について，$Q \cup \{v\}$ もまたクリークとなることに注意する．

3 形式概念

形式概念解析 (*Formal Concept Analysis*) [1] は，個体集合間の意味的な構造を解析する枠組のひとつである．

個体 (*object*) の集合 $\mathcal{O}$，および，**属性** (*feature*) の集合 $\mathcal{F}$ に対して，関係 $R \subseteq \mathcal{O} \times \mathcal{F}$ を考える．この時，タプル $\langle \mathcal{O}, \mathcal{F}, R \rangle$ を，**形式文脈** (*Formal Context*) と呼ぶ．$(o, f) \in R$ の時，個体 o は属性 f を有すると言う．個体 o が有する属性の集合 $\{f \in \mathcal{F} \mid (o, f) \in R\}$ を，$F(o)$ で参照する．

形式文脈 $\langle \mathcal{O}, \mathcal{F}, R \rangle$ に関して，写像 $\varphi : 2^{\mathcal{O}} \to 2^{\mathcal{F}}$ および $\psi : 2^{\mathcal{F}} \to 2^{\mathcal{O}}$ を考える．ここで，個体集合 $O \subseteq \mathcal{O}$ と属性集合 $F \subseteq \mathcal{F}$ について，

$$\varphi(O) = \{f \in \mathcal{F} \mid \forall o \in O, \quad f \in F(o)\} = \bigcap_{o \in O} F(o),$$

$$\psi(F) = \{o \in \mathcal{O} \mid F \subseteq F(o)\}$$

とする．つまり，φ は O 中のすべての個体が共有する属性集合を，一方，ψ は F 中のすべての属性を有する個体集合を返す写像である．

これら写像のもと，個体集合 $O \subseteq \mathcal{O}$ と属性集合 $F \subseteq \mathcal{F}$ について，$\varphi(O) = F$ かつ $\psi(F) = O$ が成り立つ時，O と F の組 $FC = (O, F)$ を**形式概念** (*Formal Concept*) [1] と定める．ここで，O と F をそれぞれ FC の**外延** (*extent*)，および，**内包** (*intent*) と呼ぶ．φ と ψ の定義より，$\psi(\varphi(O)) = O$ かつ $\varphi(\psi(F)) = F$ であることは明らかである．すなわち，形式概念とは，写像 φ と ψ に関して**閉じた** (*closed*) 個体集合 O と属性集合 F の組で与えられる．O は，F 中のすべての属性を有する個体のみから成り，かつ，それら以外にこうした個体は存在しない．同様に，F は，O 中のすべての個体に含まれる (共有される) 属性のみから成り，かつ，それら以外にこうした属性は存在しない．

形式概念 $FC = (O, F)$ および $FC' = (O', F')$ について，$O \subseteq O'$ $(F \supseteq F')$ である時，かつ，その時に限り FC と FC' 間に順序関係があるとし，これを $FC \preceq FC'$ と表記する．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．

4 核を考慮した疑似形式概念

形式概念 (O, F) は，その定義から，属性集合 F が個体集合 O を特徴付けるために必要十分であることを意味している．この様に，形式概念は外延と内包により厳密に定義されるが，その厳密さ故に，外延に大きな違いが見られない複数の異なる形式概念がしばしば観測される．概念間のこうした厳密な区別が必要である場合もあるが，一方で，データを概観する立場からは，これらを敢えて区別せずにひとつの概念として近似的に捉える方が望ましいであろう．例えば，O 中の大部分の個体が，ある属性 $f \notin F$ を有する場合，O を $F \cup \{f\}$ によって特徴付けてもさほど大きな支障はなく，我々人間の概念の捉え方と照らすとむしろ自然であると思われる．本節では，この様な柔軟な概念の捉え方を実現すべく，**疑似形式概念**を導入する．

正確な議論の前に，まずいくつかの諸概念を定義する．

定義 4.1 (外延の τ-マジョリティ)
$FC = (O, F)$ および $FC' = (O', F')$ を $FC \preceq FC'$ なる形式概念，τ を $0 \leq \tau \leq 1$ なる非負の実数とする．次が成り立つ時，かつ，その時に限り，X は X' の **τ-マジョリティ** (τ-*majority*) と呼ばれる．

$$\frac{|O|}{|O'|} \geq \tau.$$

特に，τ **マジョリティ閾値**と呼ぶ． ■

定義 4.2 (形式概念の τ-孤立性)
τ をマジョリティ閾値，$FC = (O, F) \in \mathcal{FC}$ を形式概念とする．任意の形式概念 $FC' = (O', F') \in \mathcal{FC} \backslash \{FC\}$ について，O が O' の τ-マジョリティでない時，FC は **τ-孤立** (τ-*isolated*) であると言う． ■

定義より，τ-孤立な形式概念 $FC = (O, F)$ の外延は，任意の形式概念の外延の τ-マジョリティではない．その意味で，他の形式概念に近似されるべきものではないと考えられる．よって，ここではこうした τ-孤立な形式概念を，その τ-マジョリティである外延を有する複数の形式概念の代表元と見做し，それを **τ-基底形式概念** (τ-*Base Formal Concept*) と呼ぶ．

一般に，ある概念に属するすべての個体に共通する属性は，その概念を特徴付ける**主要な属性**と考えられる．一方，すべての個体には共有されないが，その大部分の個体が有する属性もまた，概念の特徴付けにおいては，それなりの重要性を持つと考えるのが自然であろう．以下では，こうした自然な様子が反映された疑似形式概念を定義する．

定義 4.3 (τ-副次属性)
τ をマジョリティ閾値，$FC = (O, F) \in \mathcal{FC}$ を形式文脈 $\langle \mathcal{O}, \mathcal{F}, R \rangle$ における形式概念とする．属性 $f \in \mathcal{F} \backslash F$ について，$\psi(F \cup \{f\})$ が X の τ-マジョリティである時，かつ，その時に限り，f を FC の **τ-副次属性** (τ-*secondary feature*) と呼ぶ．FC の τ-副次属性の集合を $secondary_\tau(FC)$ で参照する． ■

定義 4.4 (τ-疑似形式概念)
$FC = (O, F)$ を τ-基底形式概念とする．この時，次のタプルを **τ-疑似形式概念** (τ-Pseudo Formal Concept) と呼ぶ．

$$PFC = (O, F \cup secondary_\tau(FC)).$$

ここで，O を PFC の**外延**，$F \cup secondary_\tau(FC)$ を PFC の**疑似内包**と呼ぶ．特に，F を**主要属性集合** (*Primary Feature Set*)，$secondary_\tau(FC)$ を**副次属性集合** (*Seconary Feature Set*) と呼び，F は内包の**核** (*Core*) とも呼ばれる． ■

疑似形式概念の定義から，以下が容易にわかる．

- O 中のすべての個体は，核 (主要属性集合) F を共有する．
- 任意の属性 $f \in secondary_\tau(FC)$ について，O 中の少なくとも $\tau \times 100\%$ の個体は f を有する．

5 Top-N 疑似形式概念問題

基底形式概念は，ある特定の条件を満たす形式概念であり，それをもとに疑似形式概念が (一意に) 構成されることから，疑似形式概念の総数は，形式概念のそれと比較して十分小さいことが期待できる．しかし，なおも，それらすべてを人手で解析することは現実的には不可能であると予想されることから，ここでも文献 [3, 4] 等と同様，内包制約のもとで外延評価値が上位 N である疑似形式概念のピンポイント抽出を試みる．

形式概念の外延は，内包を構成する属性群の共有を根拠とした個体の集まりであるから，属性数の小さな内包を有する外延に含まれる個体間の類似性は弱いと考えられる．外延にまとまりとしての意味を見出すために，内包に何らかの制約を課すことは極めて自然であろう．これまでの研究 [3, 4] では，抽出すべき形式概念の内包に関する制約として，それを構成する最小属性数 δ を与えた．疑似形式概念の抽出においても，これと同様の制約を課すものとする．さらに，ここでは，内包を構成する属性間の相関を考慮した制約を加えることで，概念の意味付けをより明確なものにしたい．

定義 5.1 (属性間の相関)
$\langle \mathcal{O}, \mathcal{F}, R \rangle$ を形式文脈とする．属性 $f, f' \in \mathcal{F}$ について，f と f' の**相関** $correl(f, f')$ を次の通り定める．

$$correl(f, f') = \frac{|\psi(\{f\}) \cap \psi(\{f'\})|}{|\psi(\{f\}) \cup \psi(\{f'\})|}.$$

■

定義 5.2 (属性集合の結合度)
属性集合 $F \subseteq \mathcal{F}$ について，F の**結合度** $unity(F)$ を次で定める．

$$unity(F) = \min_{f, f' \in F} \{correl(f, f')\}.$$

■

ここでは疑似内包の核に対する制約として，結合度の下限値を与える．つまり，互いに一定以上の相関を有する属性群から構成される核を有する疑似形式概念を意味あるものと考える．

以上より，本稿で解くべき問題を次の通り定める．

定義 5.3 (Top-N (τ, δ, ρ)-疑似形式概念問題)
$\langle \mathcal{O}, \mathcal{F}, R \rangle$ を形式文脈，τ $(0 \leq \tau \leq 1)$ をマジョリティ閾値，δ $(0 \leq \delta \leq |\mathcal{F}|)$ を核サイズ閾値，ρ $(0 \leq \rho \leq 1)$ を核結合度閾値，N $(N \geq 1)$ を正整数とする．この時，次を満たす τ-基底形式概念 $FC = (O, F)$ から構成される τ-疑似形式概念 $PFC = (O, F \cup secondary_\tau(FC))$ を求める問題を，*Top-N* (τ, δ, ρ)-**疑似形式概念問題**と呼ぶ．

核制約： $|F| \geq \delta$ かつ $unity(F) \geq \rho$.
目的関数： $|O|$ が上位 N.

■

6 Top-N 疑似形式概念抽出アルゴリズム概略

所与の形式文脈 $\langle \mathcal{O}, \mathcal{F}, R \rangle$，マジョリティ閾値 τ，核サイズ閾値 δ，および核結合度閾値 ρ のもとで Top-N τ-疑似形式概念を得るには，核となる内包が

制約を満たし，かつ，外延サイズが上位 N となる τ-基底形式概念を抽出すればよい．これは，重み付き無向グラフにおける**分枝限定深さ優先クリーク探索** [2] の拡張として実現される．

ここでは，ふたつの無向グラフ $G_E = (\mathcal{O}, V_E)$ と $G_I = (\mathcal{F}, V_I)$ を利用する．ここで，V_E，V_I は

$$V_E = \{(o, o') | o, o' \in \mathcal{O} \text{ and } |\varphi(\{o\}) \cap \varphi(\{o'\})| \geq \delta\},$$

$$V_I = \{(f, f') | f, f' \in \mathcal{F} \text{ and } correl(f, f') \geq \rho\}$$

である．この時，形式概念の外延は G_E におけるクリークを，結合度に関する制約を満たす内包は G_I におけるクリークをそれぞれ形成することに注意する．

探索過程においては，その時点までに抽出された Top-N 形式概念を**暫定的に保持**するリストを管理する．以下ではこれを，**暫定** *Top-N* **リスト**と呼ぶことにする．

クリークを形成する個体集合 Q について，まず最初に，その共有属性集合 $\varphi(Q)$ を求める．もし，$\varphi(Q)$ が内包のサイズ制約を満たし，かつ，クリークを形成する場合は，$\varphi(Q)$ は制約を満たす形式概念の内包となり，その外延は $\psi(\varphi(Q))$ で与えられる．こうして得られる形式概念 $(\psi(\varphi(Q)), \varphi(Q))$ が，τ-基底形式概念であれば，暫定 Top-N リストを適当に更新した後，ある拡張可能頂点 v を用いて Q を拡張し，クリーク $Q \cup \{v\}$ に対して同様の処理を再帰的に繰り返す．

一方，$\varphi(Q)$ がサイズ制約を満たさない場合は，Q の任意の拡張 Q' について，$\varphi(Q')$ もまたサイズ制約を満たさないことがわかる．よって，この場合は，Q の任意の拡張処理を安全に枝刈ることができる．$\varphi(Q)$ がクリークでない場合，$\varphi(Q)$ 中のクリークサイズの上限値を**逐次近似彩色** [2] により見積り，その値が δ に満たない場合は，Q の任意の拡張 Q' について，$\varphi(Q')$ もまたサイズ制約を満たさないことから，Q の拡張処理を同様に枝刈りできる．それ以外の場合は，そのまま Q の拡張処理を続行する．

クリーク Q を空集合で初期化した後，生成すべきクリークが存在しなくなるまで，以上の処理を深さ優先で繰り返すことで，内包制約を満たす Top-N 形式概念を漏れ無く抽出することができる．

上記した通り，探索過程において内包のサイズ制約に基づく枝刈りが可能であるが，暫定 Top-N リストに基づく枝刈りも利用可能である．クリークである個体集合 Q の拡張可能頂点集合 $cand(Q)$ を逐次近似彩色することで，Q を拡張して得られるクリークの最大値を見積もることができる．その値が，暫定 Top-N リスト中の最小外延サイズに満たない場合は，Q の任意の拡張によって Top-N 形式概念が得られないことがわかるため，この場合も Q の任意の拡張処理を安全に枝刈ることができる．

7 おわりに

本稿では，Top-N 形式概念に基づくクラスタ抽出において生ずるクラスタの重複問題に対処すべく，疑似形式概念について考察した．疑似形式概念は，外延の違いが許容誤差以内の形式概念群の代表元と捉えられる．特に，その内包は，すべての個体が共有する主要属性集合 (核) と，大部分の個体が有する副次属性集合に分割され，これら区別によって，より自然な外延の特徴付けが可能となる．所与の閾値パラメータのもと，Top-N 疑似形式概念は，分枝限定深さ優先探索により抽出可能である．今後はシステムの実装を行ない，計算機実験を通して，重複問題に対する疑似形式概念の実際の効果の確認を急ぐ．

参考文献

[1] B. Ganter and R. Wille, "Formal Concept Analysis: Mathematical Foundations", Springer, 1999.

[2] E. Tomita and T. Seki, "An Efficient Branch-and-Bound Algorithm for Finding a Maximum Clique", Proceedings of the 4th International Conference on Discrete Mathematics and Theoretical Computer Science - DMTCS'03, Springer-LNCS 2731, pp. 278 - 289, 2003.

[3] M. Haraguchi and Y. Okubo, "An Extended Branch-and-Bound Search Algorithm for Finding Top-N Formal Concepts of Documents", New Frontiers in Artificial Intelligence, JSAI 2006 Conference and Workshops, Tokyo, Japan, June 5-9, 2006, Revised Selected Papers, Springer-LNCS 4384, pp. 276 - 288, 2007.

[原口 02] 原口 誠：最適クリーク探索に基づくデータからの概念学習，人工知能学会研究会資料，SIG-FAI-A202, pp. 63 - 66, 2002.

[4] Y. Okubo and M. Haraguchi, "Finding Conceptual Document Clusters with Improved Top-N Formal Concept Search", Proc. of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence - WI'06, pp. 347 - 351, 2006.

[5] Y. Okubo, M. Haraguchi and B. Shi, "Finding Significant Web Pages with Lower Ranks by Pseudo-Clique Search", Proc. of the 8th International Conference on Discovery Science - DS'05, Springer-LNAI 3735, pp. 346 - 353, 2005.