Vol.2011-MPS-82 No.12
2011/3/7

接続概念間の構造制約に基づくレア概念抽出

大久保　好章[†1]　　原　口　　誠[†1]

本稿では，より大きな外延を有するふたつの概念を接続する示唆的概念の抽出問題について議論する．示唆的概念とは，一般的で，かつ，相関の低い属性から構成される内包を有する概念であるが，本稿では，概念間の構造に関する制約をさらに考慮することで，より興味深い示唆的概念の抽出を目指す．具体的には，抽出すべき示唆的概念は，概念的に十分離れたふたつの概念を内包的に接続するものであるとし，これにより，意外性のある隠れた関係の検出を期待する．

内包の一般性と客観的根拠に基づく単調な評価関数のもと，構造制約を満たすふたつの概念を接続し，かつ，評価値が上位 N である示唆的概念を抽出する分枝限定深さ優先探索アルゴリズムを与える．予備実験により，示唆的概念は，一見すると離れた概念間の意外な関係を明らかにできることを確認する．

Finding Indicative Concepts Connecting Larger Concepts Based on Structural Constraints

Yoshiaki OKUBO [†1] and Makoto HARAGUCHI[†1]

In this paper, we present an algorithm for finding indicative concepts with small extents connecting two concepts with larger extents, based on a structural constraint. An indicative concept has been defined as an intent of un-correlated attributes with higher supports. By the un-correlatedness, the indicative concept has smaller extent. We propose in this paper to use structural constraint as a kind of structural interestingness to restrict possible indicative concepts. Any indicative concepts under the constraint must be a common superconcept of some subconcepts of another two concepts with larger extents. As there exist many ways to have those two concepts with larger extents, we impose additional constraints to them. That is, their conceptual clarity and farness among them. Intuitively, as they are more far away, an indicative concept connecting them will be more interesting. As is actually hard to enumerate all the possible solutions satisfying the above constraints, we present a branch-and-bound procedure for investigating only top N solutions under some monotonically increasing evaluation function. We also present some experiments reporting that the procedure works well.

1. はじめに

データマイニング研究の主要なテーマのひとつである**飽和アイテム集合**[2]，あるいは，それと等価な**形式概念**[1] の抽出・列挙問題では，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する．これらと対照的に，著者らは，『**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を，**簡潔な概念** (*Concise Concepts*)[4] あるいは**示唆的概念** (*Indicative Concepts*)[5] と定め，その抽出を試みている．

本稿では，後者の示唆的概念を，それに接続する概念間の構造に関する制約を課すことで，より興味深い示唆を与え得るものへと改良する．示唆的概念は，内包の**非相関性**，**一般性**，および，**客観的根拠**を考慮して定義される[5]．具体的には，非相関性制約を満たすものの中で，一般性と客観的根拠に基づく目的関数の値が上位 N であるものを，意味解釈が容易な示唆的概念として抽出する．

本稿では，こうした示唆的概念に対して，他の概念を内包的に接続する役割を与え，接続概念間の隠れた関係や関連を見出すことを試みる．ある示唆的概念が接続可能な概念の組み合わせは，一般に多数存在することから，ここでは，接続概念の外延を構成する個体間の**概念的類似性**と**接続概念間の距離**に基づく**構造制約**により，接続可能な概念の組み合わせを有意味なものに絞り込む．概念的に十分離れた概念を接続可能な示唆的概念は，より意外性の高い概念間の関係を明らかにするものと考えられ，計算機実験によってそれを確かめる．

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．$(x, a) \in I$ の時，個体 x は属性 a を有すると言う．

形式文脈 (G, M, I) における個体集合 $X \subseteq \mathcal{G}$ と属性集合 $A \subseteq \mathcal{M}$ について，次の写像 $\varphi : 2^G \to 2^M$ と $\psi : 2^M \to 2^G$ を考える．

$$\varphi(X) = \{a \in M \mid \forall x \in X, (x, a) \in I\} \quad \text{および} \quad \psi(A) = \{x \in G \mid \forall a \in A, (x, a) \in I\}.$$

†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

これら写像のもと，個体集合 $X \subseteq G$ と属性集合 $A \subseteq M$ について，$\varphi(X) = A$ かつ $\psi(A) = X$ が成り立つ時，X と A の組 $C = (X, A)$ を**形式概念** (*Formal Concept*) [1] と定める．ここで，X と A をそれぞれ C の**外延** (*Extent*)，および，**内包** (*Intent*) と呼ぶ．以下の議論では，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 $C = (X, A)$ および $C' = (X', A')$ について，$X \subseteq X'$ $(A \supseteq A')$ 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, A) の内包 A は**飽和アイテム集合** (*Closed Itemset*)[2] に，また，外延 X は (飽和) アイテム集合 A 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 A の**頻度** ($freq(A)$ で参照) に，また，$|X|/|G|$ は A の**支持度**に一致する．

3. 非相関な一般的内包を有する示唆的概念

本節では，**示唆的概念** (*Indicative Concept*)[5] に関する諸定義を与える．示唆的概念は，文献[4] で議論された**簡潔なレア概念**の改良版であり，本稿で抽出を試みる概念の候補となる．

示唆的概念は，内包の **非相関性** および **一般性**を用いて定義される．非相関性により，それら属性の組み合わせは一般にあまり観測されないことから，こうした内包を有する概念は稀なものとなる傾向がある．また，一般性より，その概念は明快に理解可能なものであることも期待できる．

3.1 内包の非相関性

概念の内包の相関性は，*Jaccard* **係数**の拡張である *Bond* **尺度**[3] により評価する．

定義 3.1 (Bond に基づく内包の相関性)
概念 (X, A) について，その内包の相関性を $correl(A)$ と表記し，次の通り定義する．

$$correl(A) = \frac{|\bigcap_{a \in A} \psi(\{a\})|}{|\bigcup_{a \in A} \psi(\{a\})|}. \quad \blacksquare$$

一般に，概念 (X, A) について，$B \subseteq A$ かつ $\psi(A) = \psi(B) = X$ なる B が存在し，これを A の**生成元** (*generator*) と呼ぶ別の言い方をすると，X を同定するにあたり，$A \backslash B$ 中の属性は**冗長**であると言える．定義より，$correl(B) \geq correl(A)$ が成り立つことから，内包の相関性をより正確に評価するためには，こうした冗長な属性を除去することが望まれる．しかし，集合の包含関係のもとで，A の極小生成元は複数存在するため，それら極小元をすべて計算することは手間が掛かる．よって，ここではそれらの代表元を次の通り定義する．

定義 3.2 (代表生成元)
属性集合 $A = \{a_1, \ldots, a_k\}$ を考え，特に，a_i は頻度昇順にソーティングされているものとする．この時，A の**代表生成元**を $generator(A)$ と表記し，次の通り定義する．

$$generator(A) = \{a_i \in A | \psi(\{a_i\}) \not\supseteq \psi(\{a_{i+1}, \ldots, a_k\})\}. \quad \blacksquare$$

$generator(A)$ は，より低頻度の属性の組み合わせに含意されない属性の集合である．直感的に述べると，高頻度の属性は Bond 値を低下させる傾向にあることから，高頻度属性が他のいくつかの属性に含意される場合は，それを削除することが望ましい．上の定義はこの考えを反映したものである．ここでは，A の相関性を $correl(generator(A))$ により評価するものとする．

内包の**非相関性**を定義するために，上限閾値 σ による相関性制約を課す．すなわち，概念 (X, A) について，$correl(generator(A)) \leq \sigma$ が成り立つ時，内包 A は非相関であると考え，これを示唆的概念の候補として扱う．

この様に，非相関性に基づいて注目すべき概念を制限することができるが，それでもなおその数は膨大であることが文献[5] で報告されている．例えば，30,085 の語彙 (属性) を有する 2,343 の新聞記事 (個体群) からなるデータ (形式文脈) において，種々の閾値 σ のもとで，数百万の概念がしばしば観測される．非相関性を満たす概念をすべて抽出することは，一般には非現実的かつ困難なため，抽出すべき概念のさらなる質的改善が望まれる．

3.2 内包の一般性

概念の意味解釈は，その内包に基づいてなされる．よって，概念が特殊 (specific) な属性から構成される内包を有する場合，その適切な意味を与えることは困難である．そこで，容易に理解可能な概念を得るために，内包の**一般性**を考慮する．ここでは，こうした一般性を，それを構成する個々の属性の最小頻度で測るものとする．

定義 3.3 (内包の一般性)
概念 (X, A) の内包 A の**一般性**を $generality(A)$ と表記し，次の通り定義する．

$$generality(A) = \min_{a \in A}\{|\psi(\{a\})|\}. \quad \blacksquare$$

内包が頻度の高い属性だけで構成される時，その一般性は高いと考える．こうした内包は

明快に解釈可能なことが期待できる．

3.3 内包の客観的根拠

示唆的概念は内包の非相関性と一般性により定義されるが，一般に，非相関な内包に対応する外延は小さなものとなり，時には極めて少数の個体のみで構成されることもある．しかし，この様な概念は客観的な根拠に乏しく，単なる例外であるかもしれない．よって，内包の**客観的根拠**として外延の大きさを考慮するものとする．

定義 3.4 (内包の客観的根拠)
概念 (X, A) の内包 A の**客観的根拠**を $evidence(A)$ と表記し，次の通り定義する．

$$evidence(A) = |X|. \qquad \blacksquare$$

3.4 Top-N 示唆的概念抽出問題

上述の議論より，望ましい概念とは，その内包が非相関かつ一般的であり，さらに十分な客観的根拠を有するものと考え，これを**示唆的概念**と呼ぶ．

形式概念数は一般に極めて莫大であるが故に，示唆的概念もまた膨大数存在することが想像される．よって，合理的かつ現実的なアプローチとして，Top-N 法が妥当であろう．すなわち，評価値が上位 N である示唆的概念の抽出を試みる．

言うまでもなく，示唆的概念は様々な視点からの評価が可能であるが，ここでは，理解可能性と客観的根拠を重要視する立場のもと，ある度合いの非相関性を満たすものの中で，できるだけ高い一般性と客観的根拠を有する示唆的概念を望ましいものと考える．以上より，ここでの示唆的概念抽出問題を次の通り定める．

定義 3.5 (Top-N 示唆的概念マイニング)
(G, M, I) を形式文脈，σ を非相関性閾値，α を一般性重み，および，β を客観的根拠重みとする．*Top-N* **示唆的概念マイニング**とは，次の条件を満たす形式概念 $C = (X, A)$ を抽出する問題である．

制約：非相関性 $\quad correl(generator(A)) \leq \sigma.$

目的関数：一般性および客観的根拠 $\quad eval(A) = \alpha \cdot generality(A) + \beta \cdot evidence(A)$

が，非相関性制約を満たすものの中で上位 N である． $\blacksquare$

次節では，接続する概念の構造的な関係を考慮することで，示唆的概念をより示唆に富むものへと改良する．

4. 有意味な概念を接続する示唆的概念

示唆的概念は非相関な内包を有することから，その外延は小さくなる傾向にある．その結果として，容易に理解可能な**レア概念**の抽出が実現される．こうしたレア概念はそれのみでも有益に思えるが，ここではさらにその質的な改良を試みる．具体的には，より大きな概念をつなぐ**ブリッジ**の役目を課すことで，これら概念間の隠れた繋がりや関連を示唆可能なものへと発展させる．

示唆的概念 (X, A) において，個体ペア $x, y \in X$ を考える．いま，x と y がそれぞれ，より大きくかつ有意味な異なる概念の個体でもある時，x と y を，これら概念を接続する**インターフェースオブジェクト**と見做す．x と y を介して接続される概念が，例え概念的には異なるものであったとしても，示唆的概念はそれらを内包的に接続することから，示唆的概念は，接続概念間の隠れた関係を明らかにするものと言えるだろう．

本稿では，ふたつの概念を接続する示唆的概念を抽出ターゲットとする．接続概念が有意味であることを要請するために，**個体間の概念的類似性**を導入し，接続概念はそれぞれ，概念的に類似した個体から構成される外延を有することを制約として課す．

個体間の概念的類似性は，**個体** *Bond* **尺度**で測るものとする．これは，定義 3.1 における属性に対する Bond 尺度の個体版である．

定義 4.1 (個体間の概念的類似性)
個体集合 X の概念的類似度を $sim(X)$ と表記し，次の通り定義する．

$$sim(X) = \frac{|\bigcap_{x \in X} \varphi(\{x\})|}{|\bigcup_{x \in X} \varphi(\{x\})|}. \qquad \blacksquare$$

定義より，個体集合 X について，X 中の個体がほぼ同様の属性を有する時，X の類似度は高くなる．これら個体の内包的な違いはほんのわずかであるから，これらを概念的に類似したものと考えることは自然であろう．よって，概念 (X, A) について，その外延 X がある度合いの類似度を有する場合，その概念を有意味であるものと見做し，有意味な概念を接続する示唆的概念を抽出ターゲットとする．

示唆的概念は，有意味な概念間の隠れた関係を明らかにするものと期待されるが，特に，接続概念の外延に交わりがなく，かつ，十分離れたものである場合，その意外性や興味深さはより大きなものとなろう．よって，ここでは，接続概念間は一定以上の距離で離れていることを要請する．

概念間の距離は，外延中の個々の個体間の距離に基づいて定義される．

定義 4.2 (個体間の距離)
形式文脈 (G, M, I) において，M 中の各属性はある所与の順序 $M = \{a_1, \ldots, a_{|M|}\}$ で整列しているとする．また，各個体 $x \in G$ は，次に従って $|M|$-次元ベクトル $x = (v_1, \ldots, v_{|M|})$

で表現されるとする．

$$v_i = \begin{cases} 1/|\varphi(\{x\})|, & \text{if } a_i \in \varphi(\{x\}), \\ 0, & \text{otherwise.} \end{cases}$$

この時，X 中の個体 $x = (v_1, \ldots, v_{|M|})$，$y = (w_1, \ldots, w_{|M|})$ について，x と y 間の距離を $dist(x, y)$ と表記し，

$$dist(x, y) = \sqrt{\Sigma_{i=1}^{|M|}(v_i - w_i)^2}$$

と定義する． ■

個体間の距離に基づいて，概念間の距離を次の通り定義する．

定義 4.3 （概念間の距離）

概念 $C = (X, A)$ および $D = (Y, B)$ の距離を $dist(C, D)$ と表記し，次で与える．

$$dist(C, D) = min\{dist(x, y) \mid x \in X,\ y \in Y\}.$$ ■

接続概念間の距離制約に加え，一般性および客観的根拠も考慮する．すなわち，接続概念の内包は，十分な一般性を有する理解可能なものとする．さらに，接続概念の外延は，それらを接続する示唆的概念より，さらに多くの個体から構成されるものとする．こうした示唆的概念を抽出することで，より大きな概念間の隠れた関係を検出したい．

本稿で抽出を試みる示唆的概念は，先に提案された示唆的概念と比較して，構造的な制約が課せられている点でより限定的に思えるが，接続可能な概念ペアが一般に多数存在することから，それらをすべて列挙することは非現実的である．よって，ここでも Top-N アプローチを採用する．

定義 4.4 （有意味な概念を接続する Top-N 示唆的概念マイニング）

(G, M, I) を形式文脈，σ を非相関性閾値，τ を概念的類似性閾値，δ を概念距離閾値，α を一般性重み，および，β を客観的根拠重みとする．

有意味な概念を接続する *Top-N* **示唆的概念マイニング**とは，次の条件を満たす形式概念 $C = (X, A)$ を抽出する問題である．

制約：非相関性 $correl(generator(A)) \leq \sigma$.

制約：被接続概念 次を満たす有意味なふたつの概念 $C_L = (Y_L, B_L)$ と $C_R = (Y_R, B_R)$ が存在する．

- **インターフェースオブジェクト：** 個体 $x_L, x_R \in X$ について，$x_L \in Y_L \quad x_R \in Y_R$,
- **概念的類似性：** $sim(Y_L) \geq \tau, \quad sim(Y_R) \geq \tau$,
- **一般性：** 所与の K のもと，C_L について，$generality(B_L)$ は，概念的類似性制約を満たすものの中で上位 K である．C_R についても同様．
- **客観的根拠：** $|Y_L| \geq |X|, \quad |Y_R| \geq |X|$,
- **概念間距離：** $dist(C_L, C_R) \geq \delta$.

目的関数：一般性および客観的根拠 $eval(A) = \alpha \cdot generality(A) + \beta \cdot evidence(A)$ は，上記の非相関性制約と被接続概念制約を満たすものの中で，上位 N である． ■

5. 接続概念間の構造制約に基づく示唆的概念の Top-N 抽出

5.1 基本探索戦略

先にも触れた通り，ターゲットとなる示唆的概念は，一般に形式概念束中の下方に位置することから，ここでは，個体集合の拡張処理を基本とする**ボトムアップ戦略**を採用する．

形式文脈 (G, M, I) における各概念について，その外延と内包がそれぞれ $\psi(\varphi(X))$ および $\varphi(X)$ となる個体集合 $X \subseteq G$ が存在する．よって，任意の個体集合 $X \subseteq G$ について φ と ψ を適用することで，すべての概念を漏れなく抽出することができる．

いま，$G = \{x_1, \ldots, x_{|G|}\}$ について，全順序 $x_1 \prec \cdots \prec x_{|G|}$ を仮定し，G の任意の部分集合 X 中の要素は，この順序に従って整列しているものとする．

G の部分集合 $X_i = \{x_{i_1}, \ldots, x_{i_n}\}$ において，その第一要素 x_{i_1} を $head(X_i)$，最終要素 x_{i_n} を $tail(X_i)$ で参照する．また，その最初の k-要素 $\{x_{i_1}, \ldots, x_{i_k}\}$ を X_i の k-**接頭辞**と呼び，$prefix(X_i, k)$ で参照する．ここで，$0 \leq k \leq n$ であり，$prefix(X_i, 0) = \phi$ とする．

ここで，2^G 上の半順序 $\prec_s$ を次の通り定義する．

定義 5.1 （2^G 上の半順序）

G の部分集合 X_i と X_j を考える．いま，X_i が X_j の接頭辞である時，すなわち，$X_i = prefix(X_j, |X_i|)$ の時，X_i は X_j の**前者** であると言い，$X_i \prec_s X_j$ と表記する．もし，X_i が X_j の直接の前者である時，X_j を X_i の**子供**と呼ぶ． ■

$(2^G, \prec_s)$ は，ϕ を根とする**木**を構成し，特に**集合列挙木**と呼ばれる．ここでは，集合列挙木を**深さ優先**で探索する．探索中は，それまでに見つかった暫定的な Top-N 示唆的概念を格納したリストを管理する．個体集合 $X \subseteq G$ について，対応する概念 $C = (\psi(\varphi(X)), \varphi(X))$ を計算し，C が接続可能な概念が存在するか否かをチェックする．存在する場合は，C に関して暫定 Top-N リストを適切に更新した後，X の子供について同様の処理を再帰的に繰り返す．X が子供を持たない場合はバックトラックする．ϕ で初期化した X を起点として，調べるべき X が存在しなくなるまでこうした処理を深さ優先で繰り返す．

5.2 不要な拡張処理の枝刈り

Bond 尺度に基づく非相関性は，内包が小さくなるに従って単調に増加する．$X \subseteq Y$ なる $X, Y \subseteq G$ について，$\varphi(X) \supseteq \varphi(Y)$ であるから，$correl(\varphi(X)) \leq correl(\varphi(Y))$ が常に成立する．また，$generator(\varphi(X)) \subseteq \varphi(X)$ および $generator(\varphi(Y)) \subseteq \varphi(Y)$ である．よって，$correl(\varphi(X)) > \sigma$ ならば，$correl(generator(\varphi(X))) > \sigma$ かつ $correl(generator(\varphi(Y))) > \sigma$ となる．これより，次の枝刈り規則が利用可能であることがわかる．

枝刈り 5.1 ：個体集合 $X \subseteq G$ について，$correl(\varphi(X)) > \sigma$ ならば，X の任意の子供の生成は不要である． ■

これに加え，内包の上限評価値に基づいて不要な拡張処理を枝刈ることも可能である．

いま，$X \subseteq Y$ なる個体集合 $X, Y \in G$ を考える．$\varphi(X) \supseteq \varphi(Y)$ であるから，$genearlity(\varphi(X)) \leq generality(\varphi(Y)) \leq freq(head(\varphi(X)))$ を得る．つまり，$generality(\varphi(Y))$ の上限値が $freq(head(\varphi(X)))$ で与えられる．

さらに，$\psi(\varphi(X)) \subseteq \psi(\varphi(Y)) \subseteq X \cup G_{tail(X)\prec}$ であるから，$evidence(\varphi(Y))$ の上限値は $|X \cup G_{tail(X)\prec}|$ で与えられる．ここで，$G_{x\prec} = \{y \in G \mid x \prec y\}$ である．よって，$\varphi(Y)$ の上限評価値は，これら上限値の重み付き和として見積もることができることから，次の枝刈り規則が得られる．

枝刈り 5.2 ：暫定 Top-N リストにおける N 番目の評価値を min とする．個体集合 $X \subseteq G$ について，

$$\alpha \cdot freq(head(\varphi(X))) + \beta \cdot |X \cup G_{tail(X)\prec}| < min$$

ならば，X の任意の子供の生成は不要である． ■

5.3 概念の重複生成回避

一般に，ひとつの概念は複数の個体集合から計算できることから，効率的な探索を実現する上で，同一概念の重複生成回避が不可欠である．それは次の観察に基づいて実現可能である．

観察 5.1 ：個体集合 $X \subseteq G$ を考える．任意の個体 $\alpha \in \psi(\varphi(X))\backslash X$ について，$\psi(\varphi((X \cup \{\alpha\}))) = \psi(\varphi(X))$ である．すなわち，対応する概念は一致する． ■

観察 5.2 ：個体集合 $X \subseteq G$ と，その子供 $X \cup \{\alpha\}$ を考える．$\beta \prec \alpha$ なる任意の個体 $\beta \in \psi(\varphi(X \cup \{\alpha\}))\backslash\psi(\varphi(X))$ について，次を満たすある $Y \subseteq G$ が存在する．$\psi(\varphi(Y)) = \psi(\varphi((X \cup \{\alpha\})))$，かつ，$Y$ は深さ優先探索において，$X \cup \{\alpha\}$ に先行して処理済みである． ■

これら観察より，概念の重複生成を回避可能な次の枝刈り規則を得ることができる．

枝刈り 5.3 ：個体集合 $X \subseteq G$ を考える．$tail(X) \prec \alpha$ なる任意の個体 $\alpha \in \psi(\varphi(X))\backslash X$ について，$X \cup \{\alpha\}$ の生成は不要である． ■

枝刈り 5.4 ：個体集合 $X \subseteq G$ と，その子供 $X \cup \{\alpha\}$ を考える．$\beta \prec \alpha$ なる個体 $\beta \in \psi(\varphi((X \cup \{\alpha\})))\backslash\psi(\varphi(X))$ が存在するならば，$X \cup \{\alpha\}$ の生成は不要である． ■

これら枝刈り規則により，概念の重複生成の完全かつ健全な回避が可能となる．

5.4 接続概念の探索

非相関性制約を満たす概念 $C = (X, A)$ が得られる度に，構造制約のもとで C に接続可能な概念の組を探索する．構造制約を満たす組み合わせに限定しても，接続可能な概念の組み合わせは一般に多数存在するため，ここでは，各個体 $x \in X$ について，それを含む概念を Top-K 探索し，インターフェースポイントの候補 $x, y \in X$ $(x \neq y)$ について，それぞれを含む Top-K 概念同士の組み合わせのみを限定的に考えるものとする．

6. 実　　験

本節では実験の結果について述べる．実験システムは C 言語で実装し，Dual-Core AMD Opteron Processor 2222 SE(主記憶 32GB) の環境で実行した．

6.1 データセット

実験では，1994 年の毎日新聞記事をもとに，”**神戸**” が出現する 2,343 の各記事を個体とし，30,085 の中頻度名詞を属性とするデータセット (形式文脈) を作成した．このデータを DB-1994 とする．

6.2 接続概念を有する示唆的概念の例

パラメータ $\sigma = 0.01$，$\tau = 0.002$，$\delta = 0.125$，$\alpha = 1.0$ および $\beta = 1.0$ のもとで得られた接続概念を有する示唆的概念の例を以下に示す．ここで，各概念の外延中の整数は，記事の識別番号 (ID) であり，インターフェースオブジェクトには下線を付している．

示唆的概念

外延：252, 456, 707, 789, 814, 851, 1131, 1248, 1757, 1772, 2135

内包：東京，日本，大会，試合，代表，選手，チーム

接続概念 1

Extent：203, 204, 308, 327, 789, 1083, 1184, 1334, 1335, 1468, 1937, 2003, 2025, 2026, 2110

Intent：神戸，大阪，入管

接続概念 2

Extent：45, 194, 491, 774, 851, 926, 927, 1132, 1307, 1411,
1428, 1480, 1530

Intent：サッカー，川崎，鹿島

示唆的概念の外延は 11 の記事から構成され，内包の一般性は 223 であった (支持度 9.5% に相当)．外延中の記事の主な内容は次の通りである．

示唆的概念：

- 日本野球連盟発表の年度事業計画について．
- 都市対抗野球大会・予選日程について．
- 日本高校野球連盟主催，選抜高校野球大会組合せ抽選会の結果について．
- 某国代表有名サッカー選手来日時の入国拒否報道．
- サッカー・キリン杯情報．
- ラグビー・W 杯日本代表壮行試合について．
- スポーツカレンダー．
- 毎日スポーツ人賞 (毎日新聞社主催) 候補者について．

接続概念の外延はそれぞれ 15 および 13 の記事から構成され，それぞれの主な内容は次の通りである．

接続概念 1：

- 虚偽の上陸許可証印交付による某入国管理局職員の逮捕報道．
- 某国代表有名サッカー選手来日時の入国拒否報道．
- 予算成立の遅れに伴う某入国管理局業務の支障について．
- 出入国管理および難民認定法違反に問われた被告に対する判決．
- 某入国管理局支局の新運用体制について．
- 某入国管理局発表のお盆期間の出国者数予想．
- 虚偽移転による在留ビザの不法取得の実情報道．
- 外国人による虚偽住所移転問題について．

接続概念 2：

- 各種スポーツの週刊カレンダー．
- サッカー・キリン杯情報．
- J リーグの試合予定．
- J リーグの試合結果．
- 日本サッカー協会によるサッカー日本代表選手の発表．
- J リーグ・ナビスコ杯の見どころについて．

示唆的概念の記事は主に” 種々スポーツ” に関するものである．一方，接続概念はそれぞれ，” 入国管理” と” サッカー” に関係している．これらは概念的に大きく異なるものであることは容易にわかるが，インターフェースオブジェクト 789 (接続概念 1 側) と 851 (接続概念 2 側) によって接続されている．この様に，本稿での示唆的概念は，一見すると大きく離れた概念間に，こうした関連を見出す能力を有していると言えよう．示唆的概念の抽出により，隠れた意外な情報や知識に気付く機会がより多くなるものと期待している．

7. おわりに

本稿では，これまでに提案された示唆的概念をより有用にすべく，それに接続される概念間の構造制約を考慮した新たな枠組みについて議論した．特に，接続概念に対して，概念的に十分離れていることを要請することで，示唆的概念がより意外性の高い隠れた関係を明らかにすることを期待する．

内包の一般性と客観的根拠に基づく評価のもとで，評価値が上位 N である示唆的概念は，形式概念を枚挙しながら，構造制約を満たす接続概念を探索することで抽出可能である．探索過程で利用可能ないくつかの枝刈り規則についても議論した．その実装システムを用いた実験により，接続概念を考慮した示唆的概念は，一見すると離れた概念間においてさえも，隠れた意外な関係・関連を明らかにすることを確認した．

今後の課題として，接続概念の質的改良，意味を反映した概念間距離の考察，より大規模なデータに対するアルゴリズムのパフォーマンス分析等が挙げられる．

参 考 文 献

1) B. Ganter and R. Wille, Formal Concept Analysis - Mathematical Foundations, 284 pages, Springer, 1999.
2) N. Pasquier, Y. Bastide, R. Taouil and L. Lakhal, Efficient Mining of Association Rules Using Closed Itemset Lattices, Information Systems, 24(1), 25 – 46, 1999.
3) E. R. Omiecinski, Alternative Interest Measures for Mining Associations in Databases, IEEE Transactions on Knowledge and Data Engineering, 15(1), 57 – 69, 2003.
4) Y. Okubo and M. Haraguchi, An Algorithm for Extracting Rare Concepts with Concise Intents, Proc. of the 8th International Conference on Formal Concept Analysis - ICFCA’10, LNAI-5986, 145 – 160, 2010.
5) Y. Okubo, M. Haraguchi and T. Nakajima, Finding Rare Patterns with Weak Correlation Constraint, Proc. of the 2010 IEEE International Conference on Data Mining Workshops - ICDMW’10, 822 – 829, 2010.