部分的関係からのクラスタリング：直観的データ解析の枠組

斉藤康彦[†,††] 東条敏[†,*] 古宮誠一[†,†††]

実世界における知識の獲得では，観察された事例を直観的に分類しなければならないことがある．しかし，一般には，事例を特徴付ける属性が未知であることから，厳密な分類規則を得ることが難しい．そこで，本論文は，属性を用いずに記述した事例を，分類規則を用いずに分類する枠組を提案する．本枠組では，部分的な事例間の類義関係と反義関係から，一貫性のあるクラスタリングを生成する．このとき，典型的な事例の集合としてのクラスタからなる，分類のモデルを与える．事例と典型的な事例の集合の間の親和性に基づいて，暫定的なクラスタリングを生成し，漸次，これを再編していくことによって，与えられたモデルに準拠した一貫性のあるクラスタリングを導く．本枠組は，探索的データ解析の手法である．変量間の相関係数が計算できないために，因子分析が適用できないような場合にも，直観にしたがって仮説の探索が行えることを示す．

Clustering from Partial Relations: A Framework for Intuitive Data Analysis

YASUHIKO SAITO,[†,††] SATOSHI TOJO[†,*] and SEIICHI KOMIYA[†,†††]

To acquire knowledge in real-world domains, one classifies observed instances intuitively. However, it is difficult to obtain rigorous classification rules, because attributes characterizing instances are usually unknown. This paper, therefore, proposes a framework for describing instances without using any attribute and for classifying them without using any classification rule. In the framework, a coherent clustering is generated from partial relations: synonymity and antonymity between instances. In addition to partial relations between instances, a set of clusters, each of which contains typical instances, is given as a model of classification. Then, a temporary clustering is generated on the basis of affinity between each instance and each cluster containing typical instances. Next, the clustering is gradually reorganized. Thus, a coherent clustering, which conforms to the given model of classification, is derived. The framework serves as a method for exploratory data analysis. Factor analysis is not used, if correlation coefficients between variables cannot be calculated. In this case, however, using the framework, one can explore hypotheses according to intuition.

1. はじめに

概念帰納学習[6]では，訓練事例から概念記述を帰納する．概念記述は，事例の分類規則である．概念学習システムは，まず，分類規則を生成し，次いで，分類規則を用いて，新たな事例を分類する．それが可能なのは，ヴァージョン空間[7]の枠組において明らかであるように，事例を特徴付ける記述が正確で首尾一貫していること，および，概念記述が事例を特徴付ける記述から一般化されることを前提としているからである．

多くの概念学習システムでは，このような前提に基づいた表現言語によって，事例や概念が記述される．たとえば，ID3における決定木[10]や，CLUSTER/Sにおける拡張述語論理[13]などである．これらの表現言語は，論理的に厳密な分類規則の記述に適している．

これに対して，独立した知識としての分類規則を獲得することよりも，新たな事例を分類することに重点をおいた研究では，確率的な表現によって分類規則が記述されたり[2]，事例に基づく学習[1]のように，分類規則が存在しなかったりする．

これらの研究に共通することは，事前に与えられた属性を用いて事例を特徴付けていることである．分類規則を生成する場合には，属性を用いて分類規則が記述される．分類規則を生成しない場合にも，事例を特

† 情報処理振興事業協会 (IPA)
Information-technology Promotion Agency, Japan (IPA)
†† (株)アイネス
INES Corp.
††† (株)日立製作所
Hitachi Ltd.
* 現在，(株)三菱総合研究所
Presently with Mitsubishi Research Institute

徴付けている属性を用いないで，新たな事例が属するクラスを決定することはできない．

しかし，属性が事前に与えられているという想定は，必ずしも現実的でない．確かに，既存のデータベースから知識を獲得するような場合には，属性が既知であるが，実世界において観察される事例では，一般に，属性が未知である．また，イメージにしたがって何となく分類した結果から，逆に，適切な属性が見い出されるようなケースもある．

本論文では，属性に基づく分類に対して，イメージに基づく分類，すなわち，直観的分類の枠組を提案する．本枠組においては，属性を用いずに事例を記述する．したがって，分類規則を生成しない．多くの概念学習システムの目的は，分類規則を生成することであったが，本枠組の目的は，分類規則を用いずにクラスタリングを生成することである．これまで概念帰納学習の領域で対象とされてきた属性に基づく分類は，分類の一面でしかない．実世界における知識の獲得のためには，直観的分類が不可欠である．

本論文は，次のような構成になっている．第2章では，本枠組と因子分析[12]の関連を論ずる．因子分析は，統計に基づく直観的分類の手法である．第3章では，本枠組の基礎となる，部分的関係からの分類について説明する．第4章では，諸定義，分類アルゴリズムを提示し，例を用いて説明する．第5章では，本枠組によるデータ解析の実験について論ずる．ここでは，因子分析による解析と比較する．

2. 標本の分類と変量の分類

本章では，因子分析における直観的分類との関連について述べる．

因子分析は，多変量解析の一手法である．因子とは，変量間の相関を説明する，潜在的な変量のことである．まず，各標本を行とし各変量を列とするデータ行列に基づいて，変量間の相関係数を計算する．次いで，変量間の相関係数に基づいて，因子負荷量，および，因子得点を計算する．因子負荷量は，各変量と各因子の結合の強さを示す．ひとつの因子と強く結合したいくつかの変量は，ひとつのクラスタを形成する．因子得点は，各標本と各因子の結合の強さを示す．ひとつの因子と強く結合したいくつかの標本は，ひとつのクラスタを形成する．

クラスタリング手法としての因子分析には，2通りの適用法がある．因子得点にしたがって標本を分類する適用法と，因子負荷量にしたがって変量を分類する適用法である．たとえば，作曲家を分類することを考えてみる．作曲された交響曲の数，協奏曲の数，…などを変量とし，作曲家を標本とする場合には，因子得点を計算することによって，標本である作曲家を分類する．これは，属性に基づく分類である．すなわち，これらの変量は，各作曲家を特徴付ける属性である．これに対して，作曲家を変量とし，音楽学校の生徒の作曲家に対する好き嫌いを標本とする場合には，因子負荷量を計算することによって，変量である作曲家を分類する．これは，直観に基づく分類である．すなわち，これらの変量間の相関は，多くの人間の直観から導かれたものである．

後者の適用法は，属性を必要としないが，標本として，多くの人間の直観を必要とする．相関係数の微妙な差の根拠は，標本の数が多いことに依存している．また，標本に誤りがあったとしても，多くの標本の中に埋もれてしまうことを前提としている．したがって，少数の人間の直観を扱うには，この手法は適していない．標本の数が十分に多くなければ，相関係数の微妙な差は信用できない．特に，ひとりの人間の直観から，相関が生じることはない．

そこで，本論文では，ひとり，もしくは，少数のデータ解析者が，多くの標本を必要とすることなく，直観的に変量を分類することを支援する枠組を提案する．本枠組は，因子分析と同じように，分類すべき変量間の関係を解析の対象とする．しかし，解析の方法は，行列計算のような数値的な操作によらない．

3. 部分的関係と分類

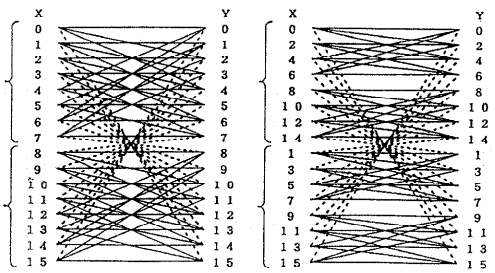
本章では，部分的にしか定義されていない不完全な分類から，首尾一貫した分類を導き出す枠組の，基本的な考え方について述べる．

因子分析における変量間の相関に対応するものとして，事例間の類義関係と反義関係を考える．これらの関係は，ふたつの事例について，

（1） 類義性がある→同じクラスタに属する

（2） 反義性がある→同じクラスタに属さない

ということを規定したものである．正の相関が類義関係に対応し，負の相関が反義関係に対応する．類義関係と反義関係は，少数の事例のみに着目して，部分的に定義した関係なので，全体としては，矛盾していることがある．たとえば，以下の例のような関係が成立するかもしれない．



X ——— Y : Y is synonymous with X.
X - - - - - Y : Y is antonymous with X.

図 1 部分的関係と分類
Fig. 1 Partial relations and classification.

[例 1] $\begin{cases} synonymous\ (\text{long, large}) \\ antonymous\ (\text{large, long}) \end{cases}$

[例 2] $\begin{cases} synonymous\ (\text{long, large}) \\ synonymous\ (\text{large, small}) \\ synonymous\ (\text{small, short}) \\ antonymous\ (\text{short, long}) \end{cases}$

したがって，部分的な関係である類義関係と反義関係は，一般に，不完全な分類である．

そこで，部分的関係における一貫性の欠如を補うために，分類の意図を暗示するモデルを，大局的な視点から与える．たとえば，例1のような場合に，"long" と "large" が同じクラスタに属する可能性は，"long" と "large" が同じクラスタに属するようなモデルを与えると，相対的に大きくなり，"long" と "large" が異なるクラスタに属するようなモデルを与えると，相対的に小さくなる．

部分的関係から分類を導き出す場合に，分類が一意的に定まらないことがある．たとえば，**図1**は，0から15までの整数における類義関係と反義関係であるが，左のダイアグラムは，7以下の整数のクラス（上半分）と8以上の整数のクラス（下半分）に分かれる．すなわち，類義性を表す実線は，同じクラスに属する要素を結んでおり，反義性を表す破線は，異なるクラスに属する要素を結んでいる．一方，部分的関係を保存したまま，これらの整数を並び換えただけの右のダイアグラムは，偶数のクラス（上半分）と奇数のクラス（下半分）に分かれる．しかし，何通りかの分類が考えられる場合にも，与えられたモデルに準拠することによって，適切な分類を導き出すことができる．

4. 親和性に基づく直観的分類

本章では，親和性に基づく直観的分類[11]の枠組について詳細に述べる．本枠組は，事例間の部分的関係と分類のモデルから，クラスタリングを生成する．まず，基本概念の定義と分類のアルゴリズムを示す．次に，説明のための簡単な例を示す．

4.1 定　　義

[類義語と反義語]

本枠組では，分類すべきそれぞれの事例を，ひとつの語とみなす．すべての語は，類義語の集合と，可能であるならば，反義語の集合をもつ．ここで，語wの類義語の集合を $w.Syn$ とし，語wの反義語の集合を $w.Ant$ とする．

[モデルクラスタリング]

モデルクラスタリングは，モデルクラスタの一組として与えられる．各モデルクラスタは，そのクラスタの典型的な語を含む．すべての語は，モデルクラスタリングにしたがって分類される．

[親和性]

語wとモデルクラスタCの間の親和性は，以下のように定義される．

$$aff(w,C)=ss(w,C)+sa(w,C)+as(w,C)+aa(w,C)$$

ただし，

$$ss(w,C)=\begin{cases} 1, & if\ w.Syn\cap\cup_{a\in C}(a.Syn)\neq\phi \\ 0, & otherwise \end{cases}$$

$$sa(w,C)=\begin{cases} -1, & if\ w.Syn\cap\cup_{a\in C}(a.Ant)\neq\phi \\ 0, & otherwise \end{cases}$$

$$as(w,C)=\begin{cases} -1, & if\ w.Ant\cap\cup_{a\in C}(a.Syn)\neq\phi \\ 0, & otherwise \end{cases}$$

$$aa(w,C)=\begin{cases} 1, & if\ w.Ant\cap\cup_{a\in C}(a.Ant)\neq\phi \\ 0, & otherwise \end{cases}$$

[最適クラスタ]

親和性に基づいて，各語の最適クラスタが決定する．最適クラスタは，モデルクラスタのひとつであり，各語は，それぞれの最適クラスタであるモデルクラスタに振り分けられる．任意の語について，最適クラスタ候補との間の親和性は，他のモデルクラスタとの間の親和性よりも大きいか等しいが，最適クラスタとの間の親和性は，他のモデルクラスタとの間の親和性よりも大きくなければならない．したがって，最適クラスタ候補は必ず存在するが，最適クラスタが決定するとは限らない．

4.2 アルゴリズム

4.2.1 概　　要

与えられたすべての語についての最適クラスタを決定するために，暫定的に生成されたクラスタリングを徐々に再編していく．本アルゴリズムは，分配モジ

ュールと変形モジュールからなる．

分配モジュールは，最適クラスタの条件を緩和しながら，順次，語を最適クラスタに分配していく．モデルクラスタ C を最適クラスタとする語の集合を $Gen(C)$ とする．語 w の最適クラスタが C であるときに，$w \in Gen(C)$ である $Gen(C)$ が生成される．最適クラスタが決定しない語は，どのモデルクラスタにも分配されずに，残存語として扱われる．

変形モジュールは，少なくともひとつの残存語についての最適クラスタが決定するように，モデルクラスタリングを変形する．最終的なクラスタリングは，与えられたモデルクラスタリングに準拠していなければならない．このために，以下の戦略にしたがう．

1. 対象とする残存語の最適クラスタは，モデルクラスタリングの変形によって必ず決定する．
2. ある語の最適クラスタがすでに決定しているならば，その語の最適クラスタはモデルクラスタリングの変形の後にも決定する（ここで，変形前の最適クラスタと変形後の最適クラスタが，同じものである必要はない）．

よって，残存語の数は，変形のたびに減少していく．

分配モジュールと変形モジュールは，残存語がなくなるまで交互に実行される．

4.2.2 分配モジュール

本モジュールでは，処理中の語の最適クラスタが決定するまで，最適クラスタの定義が拡張される．拡張された定義に適合するモデルクラスタは，必ず最適クラスタ候補である．最適クラスタの定義は，以下に記す順に拡張され，それでもなお最適クラスタが決定しない語は，残存語として扱われる．

[拡張 1]

現在処理中の語を w とする．n 個のクラスタを含むモデルクラスタリングを M とする．このとき，以下の手続きを実行した結果，$sum(C_i) > \left(\sum_{k=1}^{n} sum(C_k) \right) \Big/ 2$ であり，かつ，C_i が w の最適クラスタ候補であるならば，C_i は，w の最適クラスタである．

```
for each Ci∈M,
  sum(Ci) :=0;
  for each s∈w.Syn,
    if ∀x∈M: aff(s, Ci)≥aff(s, x), then
      sum(Ci) :=sum(Ci)+1;
  for each a∈w.Ant,
    if ∀x∈M: aff(a, Ci)≤aff(a, x), then
      sum(Ci) :=sum(Ci)+1;
```

本手続きでは，まず，現在処理中の語のすべての類義語について，親和性の値が最大のモデルクラスタを数えあげ，次に，現在処理中の語のすべての反義語について，親和性の値が最小のモデルクラスタを数えあげている．その過半数を占めるモデルクラスタが，最適クラスタ候補である場合に，その最適クラスタ候補を最適クラスタとする．拡張 1 は，特に，本モジュールを最初に実行した後に，残存語が多くなりすぎないようにすることを意図している．

[拡張 2]

現在処理中の語の，モデルクラスタリングの変形前の最適クラスタを P とする．現在処理中の語の，モデルクラスタリングの変形後の最適クラスタ候補のひとつを Q とする．このとき，$P \supseteq Q$ であるならば，Q は，現在処理中の語の最適クラスタである．拡張 2 は，特に，ある語の最適クラスタが，後述する変形モジュールにおける縮小化によって書き換えられても，その語の最適クラスタを決定できるようにすることを意図している．

[拡張 3]

現在処理中の語の，モデルクラスタリングの変形前の最適クラスタを P とする．現在処理中の語の，モデルクラスタリングの変形後の最適クラスタ候補のひとつを Q とする．このとき，$P \subseteq Q$ であるならば，Q は，現在処理中の語の最適クラスタである．拡張 3 は，特に，ある語の最適クラスタが，後述する変形モジュールにおける統合化によって書き換えられても，その語の最適クラスタを決定できるようにすることを意図している．

[拡張 4]

現在処理中の語のみを含むクラスタは，その語の最適クラスタである．拡張 4 は，後述する変形モジュールにおける孤立化によって，確実に，現在処理中の語の最適クラスタを決定できるようにすることを意図している．

4.2.3 変形モジュール

本モジュールでは，対象とする残存語の最適クラスタが決定するように，縮小化，統合化，孤立化の変形操作が，モデルクラスタリングに対して，以下に記す順に試みられる．縮小化と統合化は成功するとは限らないが，孤立化は必ず成功する．

[縮小化]

本操作は，ひとつのモデルクラスタからいくつかの語を除去する．対象とする残存語を w とする．w の最

適クラスタ候補を C とする．$\{x|w.Syn \cap x.Ant = \phi, x \in C\}$，または，$\{x|w.Ant \cap x.Syn = \phi, x \in C\}$ を R とする．以下をすべて満足するときに，C が R によって置き換えられる．

$$aff(w,R) = aff(w,C)+1$$

$$\forall y : (y \in Gen(C)) \rightarrow (aff(y,R) \geq aff(y,C))$$

[統合化]

本操作は，ふたつのモデルクラスタを併合する．対象とする残存語を w とする．w の最適クラスタ候補を C_1, C_2 とする．$C_1 \cup C_2$ を U とする．以下をすべて満足するときに，U がモデルクラスタリングに追加され，同時に，C_1, C_2 がモデルクラスタリングから削除される．

$$aff(w,U) = aff(w,C_1)+1 = aff(w,C_2)+1$$

$$\left(\begin{array}{l}\text{最適クラスタ候補がふたつだけならば，} \\ \quad aff(w,U) = aff(w,C_1)+1 = aff(w,C_2)+1 \\ \text{または，} \\ \quad aff(w,U) = aff(w,C_1) = aff(w,C_2)\end{array}\right)$$

$$\forall x : (x \in Gen(C_1)) \rightarrow (aff(x,U) \geq aff(x,C_1))$$

$$\forall x : (x \in Gen(C_2)) \rightarrow (aff(x,U) \geq aff(x,C_2))$$

[孤立化]

本操作は，対象とする残存語のみを含むクラスタを新たに生成する．

4.3 色の分類

本節では，クラスタリングの再編が不要な分類の例を用いて，分配モジュールにおける処理を詳説する．

表1は，“白”，“黒”，“灰”，“赤”，“青”，“緑”，“黄”，“朱”，“紫”，“紺”の間の類義関係と反義関係である．ここで，$w.Syn = \{s_1, s_2, \cdots, s_m\}$，および，$w.Ant = \{a_1, a_2, \cdots, a_n\}$ を，次のように表している．

$$w : (\{s_1, s_2, \cdots, s_m\}, \{a_1, a_2, \cdots, a_n\})$$

これらの色を，以下のモデルクラスタリングにしたがって分類する．

表1 色の間の部分的関係
Table 1 Partial relations between colors.

Word	Synonym	Antonym
白：({白, 朱},		{黒, 紫})
黒：({黒, 灰, 紫, 紺},		{白, 朱})
灰：({灰, 紫, 紺},		{白, 赤, 朱})
赤：({赤, 白},		{灰, 青, 緑, 黄})
青：({青, 黒, 灰, 緑, 黄},		{赤})
緑：({緑, 黄},		{赤, 青})
黄：({黄, 緑},		{赤, 青})
朱：({朱, 白},		{紫})
紫：({紫, 黒},		{朱})
紺：({紺, 黒, 灰, 紫},		{白, 朱})

C_1 = {白, 朱}，C_2 = {黒, 紫}，C_3 = {青}

分配モジュールは，各色と各モデルクラスタの間の親和性を計算することによって，各色の最適クラスタを決定する．しかし，“灰”については，次のような計算の結果，最適クラスタが決定しない．モデルクラスタリングから，

$$\cup_{a \in C_1}(a.Syn) = 白.Syn \cup 朱.Syn = \{白, 朱\}$$
$$\cup_{a \in C_1}(a.Ant) = 白.Ant \cup 朱.Ant = \{黒, 紫\}$$
$$\cup_{a \in C_2}(a.Syn) = 黒.Syn \cup 紫.Syn = \{黒, 灰, 紫, 紺\}$$
$$\cup_{a \in C_2}(a.Ant) = 黒.Ant \cup 紫.Ant = \{白, 朱\}$$
$$\cup_{a \in C_3}(a.Syn) = 青.Syn = \{黒, 灰, 青, 緑, 黄\}$$
$$\cup_{a \in C_3}(a.Ant) = 青.Ant = \{赤\}$$

C_1 に対しては，

$$灰.Syn \cap \cup_{a \in C_1}(a.Syn) = \phi$$
$$灰.Syn \cap \cup_{a \in C_1}(a.Ant) \neq \phi$$
$$灰.Ant \cap \cup_{a \in C_1}(a.Syn) \neq \phi$$
$$灰.Ant \cap \cup_{a \in C_1}(a.Ant) = \phi$$

C_2 に対しては，

$$灰.Syn \cap \cup_{a \in C_2}(a.Syn) \neq \phi$$
$$灰.Syn \cap \cup_{a \in C_2}(a.Ant) = \phi$$
$$灰.Ant \cap \cup_{a \in C_2}(a.Syn) = \phi$$
$$灰.Ant \cap \cup_{a \in C_2}(a.Ant) \neq \phi$$

C_3 に対しては，

$$灰.Syn \cap \cup_{a \in C_3}(a.Syn) \neq \phi$$
$$灰.Syn \cap \cup_{a \in C_3}(a.Ant) = \phi$$
$$灰.Ant \cap \cup_{a \in C_3}(a.Syn) = \phi$$
$$灰.Ant \cap \cup_{a \in C_3}(a.Ant) \neq \phi$$

よって，

$$aff(灰, C_1) = 0+(-1)+(-1)+0 = -2$$
$$aff(灰, C_2) = 1+0+0+1 = 2$$
$$aff(灰, C_3) = 1+0+0+1 = 2$$

したがって，C_2 と C_3 が“灰”の最適クラスタ候補であるが，“灰”の最適クラスタは決定しない．なお，“灰”以外の色についても計算すると，

	白	黒	灰	赤	青	緑	黄	朱	紫	紺
C_1	2	−2	−2	1	−1	0	0	2	−2	−2
C_2	−2	2	2	−2	1	0	0	−2	2	2
C_3	−1	1	2	−2	2	1	1	0	1	1

となるので，“灰”以外の最適クラスタは決定する．

“灰”の最適クラスタを決定するために，最適クラスタの条件を，拡張1にしたがって緩和する．分配モジュールにおいて現在処理中の語が“灰”であるとする．まず，“灰”の各類義語，すなわち，“灰”，“紫”，“紺”と各モデルクラスタの間の親和性は，すでに計

算したとおりであるから，

$sum(C_1)=0$

$sum(C_2)=3$

$sum(C_3)=1$

次に，“灰”の各反義語，すなわち，“白”，“赤”，“朱”と各モデルクラスタの間の親和性は，すでに計算したとおりであるから，

$sum(C_1)=0+0=0$

$sum(C_2)=3+3=6$

$sum(C_3)=1+1=2$

したがって，$sum(C_2)>\left(\sum_{k=1}^{3} sum(C_k)\right)\Big/2$ となり，さらに，C_2 は“灰”の最適クラスタ候補なので，C_2 が“灰”の最適クラスタである．こうして，以下のクラスタリング結果が得られる．

$Gen(C_1)=$ {白, 赤, 朱}

$Gen(C_2)=$ {黒, 灰, 紫, 紺}

$Gen(C_3)=$ {青, 緑, 黄}

4.4 作家の分類

本節では，クラスタリングの再編が必要な分類の例を用いて，変形モジュールにおける処理を詳説する．

表 2 は，適当に選んだ日本の作家 30 人の間の類義関係と反義関係である．これらの作家を，以下のモデルクラスタリングにしたがって分類する．

$C_1=$ {森鷗}

$C_2=$ {夏漱，芥竜，三由}

$C_3=$ {島藤，志直}

$C_4=$ {永荷，谷潤，横利，太治}

$C_5=$ {石淳，堀辰}

$C_6=$ {中重}

このとき，以下のクラスタリング結果が得られる．

$Gen(C_1)=$ {森鷗}

$Gen(C_2)=$ {菊寛，芥竜，三由}

$Gen(C_3)=$ {国独，田花，島藤，有武，志直，葛善}

$Gen(C_4)=$ {梶基，坂安，井靖，太治}

$Gen(C_5)=$ {井鱒，石淳，堀辰，伊整，大昇}

$Gen(C_6)=$ {中重，小多，野宏}

また，分類されない作家は，“二四”，“夏漱”，“泉鏡”，“永荷”，“武実”，“谷潤”，“横利”，“川康”の 8 人である．

再編 1：生成されたクラスタリングを，“二四”の最適クラスタを決定するために再編する．ここで，“二四”の最適クラスタ候補は C_1 と C_2 であるが，C_1 と C_2 に対して統合化が可能である（その論拠となる計算の詳細を付録Aに示す）．したがって，C_1 と C_2 の

表 2 作家の間の部分的関係
Table 2 Partial relations between novelists.

Word	Synonym	Antonym
森鷗：	({森鷗，夏漱，志直}，	{})
二四：	({二四，夏漱}，	{})
夏漱：	({夏漱，森鷗，芥竜}，	{})
国独：	({国独，田花}，	{})
田花：	({田花，国独}，	{})
島藤：	({島藤，田花，志直}，	{泉鏡})
泉鏡：	({泉鏡，谷潤}，	{二四，森鷗})
有武：	({有武，武実，志直}，	{})
永荷：	({永荷，谷潤}，	{中重})
志直：	({志直，森鷗，武実}，	{泉鏡，谷潤})
武実：	({武実，有武，志直}，	{坂安，太治})
谷潤：	({谷潤，泉鏡，川康}，	{島藤})
葛善：	({葛善，田花，太治}，	{泉鏡，谷潤})
菊寛：	({菊寛，芥竜}，	{})
芥竜：	({芥竜，谷潤，三由}，	{武実})
井鱒：	({井鱒，石淳}，	{})
横利：	({横利，川康}，	{泉鏡})
石淳：	({石淳，泉鏡，谷潤}，	{小多})
川康：	({川康，谷潤，梶基}，	{})
梶基：	({梶基，横利，川康}，	{})
中重：	({中重，小多，野宏}，	{谷潤})
小多：	({小多，中重，野宏}，	{谷潤})
堀辰：	({堀辰，大昇，伊整}，	{小多，永荷})
伊整：	({伊整，大昇，井靖}，	{})
坂安：	({坂安，太治}，	{})
井靖：	({井靖，川康}，	{三由})
太治：	({太治，葛善}，	{森鷗，志直})
大昇：	({大昇，堀辰}，	{})
野宏：	({野宏，中重}，	{})
三由：	({三由，泉鏡，芥竜，谷潤}，	{武実，中重})

森鷗＝森鷗外	二四＝二葉亭四迷	夏漱＝夏目漱石
国独＝国木田独歩	田花＝田山花袋	島藤＝島崎藤村
泉鏡＝泉鏡花	有武＝有島武郎	永荷＝永井荷風
志直＝志賀直哉	武実＝武者小路実篤	谷潤＝谷崎潤一郎
葛善＝葛西善蔵	菊寛＝菊池寛	芥竜＝芥川竜之介
井鱒＝井伏鱒二	横利＝横光利一	石淳＝石川淳
川康＝川端康成	梶基＝梶井基次郎	中重＝中野重治
小多＝小林多喜二	堀辰＝堀辰雄	伊整＝伊藤整
坂安＝坂口安吾	井靖＝井上靖	太治＝太宰治
大昇＝大岡昇平	野宏＝野間宏	三由＝三島由紀夫

合併をモデルクラスタリングに追加し，C_1 と C_2 をモデルクラスタリングから削除する．こうして，以下のモデルクラスタリングと，それにしたがったクラスタリング結果が得られる．

$C_3=$ {島藤，志直}

$C_4=$ {永荷，谷潤，横利，太治}

$C_5=$ {石淳，堀辰}

$C_6=$ {中重}

$C_7=$ {森鷗，夏漱，芥竜，三由}

$Gen(C_3)=$ {森鷗，国独，田花，島藤，有武，志直，武実，葛善}

$Gen(C_4)$= {梶基，坂安，井靖，太治}
$Gen(C_5)$= {井鱒，石淳，堀辰，伊整，大昇}
$Gen(C_6)$= {中重，小多，野宏}
$Gen(C_7)$= {二四，夏漱，菊寛，芥竜，三由}

また，分類されない作家は，“泉鏡”，“永荷”，“谷潤”，“横利”，“川康”の5人である．

再編2：生成されたクラスタリングを，“泉鏡”の最適クラスタを決定するために再編する．ここで，“泉鏡”の最適クラスタ候補は C_4 と C_5 であるが，“横利”を除去することで，C_4 に対して縮小化が可能である（その論拠となる計算の詳細を付録Bに示す）．したがって，C_4 を縮小化されたクラスタで置き換える．こうして，以下のモデルクラスタリングと，それにしたがったクラスタリング結果が得られる．

C_3= {島藤，志直}
C_4= {永荷，谷潤，太治}
C_5= {石淳，堀辰}
C_6= {中重}
C_7= {森鷗，夏漱，芥竜，三由}
$Gen(C_3)$= {森鷗，国独，田花，島藤，有武，志直，武実，葛善，横利}
$Gen(C_4)$= {泉鏡，谷潤，川康，梶基，坂安，井靖，太治，三由}
$Gen(C_5)$= {井鱒，石淳，堀辰，伊整，大昇}
$Gen(C_6)$= {中重，小多，野宏}
$Gen(C_7)$= {二四，夏漱，菊寛，芥竜}

また，分類されない作家は，“永荷”だけである．

再編3：生成されたクラスタリングを，“永荷”の最適クラスタを決定するために再編する．ここで，“永荷”の最適クラスタ候補は C_4 と C_7 であるが，縮小化も統合化も可能でない（その論拠となる計算の詳細を付録Cに示す）．したがって，{永荷} をモデルクラスタリングに追加する．こうして，以下のモデルクラスタリングと，それにしたがったクラスタリング結果が得られる．

C_3= {島藤，志直}
C_4= {永荷，谷潤，太治}
C_5= {石淳，堀辰}
C_6= {中重}
C_7= {森鷗，夏漱，芥竜，三由}
C_8= {永荷}
$Gen(C_3)$= {森鷗，国独，田花，島藤，有武，志直，武実，葛善，横利}
$Gen(C_4)$= {泉鏡，谷潤，川康，梶基，坂安，井靖，太治，三由}
$Gen(C_5)$= {井鱒，石淳，堀辰，伊整，大昇}
$Gen(C_6)$= {中重，小多，野宏}
$Gen(C_7)$= {二四，夏漱，菊寛，芥竜}
$Gen(C_8)$= {永荷}

これによって，すべての作家が分類されたので，アルゴリズムは停止する．

5. データ解析の実験

本章では，親和性に基づく直観的分類の枠組によるデータ解析の実験について述べる．さらに，現実的なデータ解析に適用する場合の留意点について考察する．

本枠組と因子分析は，事例（変量）間の関係からクラスタリングを生成する．しかし，本枠組は，変量間の相関係数が計算できない場合にも，適用することができる．すなわち，標本の数が少なかったり，もともと直観的にしか相関が与えられないような場合であ

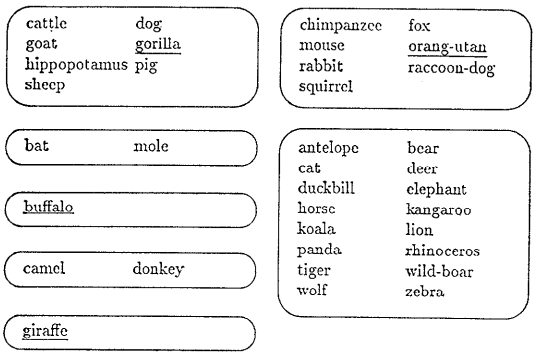


図2 動物の直観的分類（修正前）
Fig. 2 An intuitive classification of animals (the first trial).

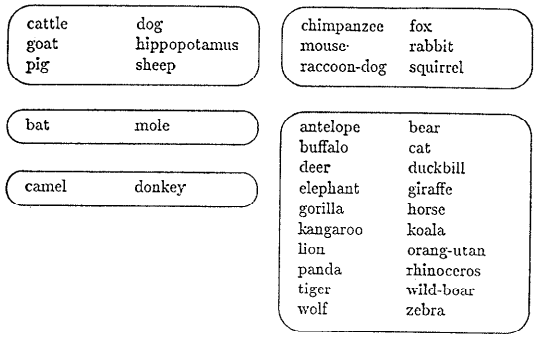


図3 動物の直観的分類（修正後）
Fig. 3 An intuitive classification of animals (the last trial).

る．このような場合のデータ解析として，36種類の動物を分類する実験を行った．事例間の部分的関係とモデルクラスタリングは，客観的な分類の基準を設けずに，直観的に与えた．**図2**は，このとき最初に得られたクラスタリングであるが，下線で示した動物の分類が，何となく不適切であると感じた．そこで，クラスタリングの再編の過程，すなわち，与えられたモデルクラスタリングと変形後のモデルクラスタリングの差分を参考にしながら，事例間の部分的関係とモデルクラスタリングを修正した．これは，直観に誤りがあると判断したからである．**図3**は，このような修正の後に得られたクラスタリングである．このクラスタリングは，納得のいく若干の再編を経て得られる．

変量間の相関係数が計算できる場合には，因子分析による解析結果と本枠組による解析結果を比較することができる．そこで，次のような実験を行った．

1. 150人を対象に，各動物に対する好感度を5段階で評価するようなアンケート調査を行い，データ行列を作成した．各動物が変量であり，各回答者が標本である．
2. データ行列に基づいて，相関行列（変量間の相関の行列）を作成した．
3. 相関行列に基づいて，因子行列（各変量と各因子の間の相関の行列）を作成した．
4. 因子行列に基づいて，分類を行った．**図4**は，因子分析によるクラスタリング結果である．
5. 相関行列において，相関係数が0.4以上の相関に対して類義性を，相関係数が-0.1以下の相関に対して反義性を定義した．**図5**は，このようにして得られた行列である．なお，ここで設定した境界値に特別な根拠はないので，適当に変更して実験を行うことができる．
6. 暫定的なモデルクラスタリングを与えて，分類を行った．
7. 相関行列から得られた事例間の部分的関係に対して，モデルクラスタリングが整合的でなかったために，クラスタリングの再編の過程は，簡潔なものにならなかった．そこで，このときの再編の過程を参考にしながら，モデルクラスタリングを修正した．
8. クラスタリングの再編が不要になるまで，ステップ6，7を繰り返し，最終的に，以下のようなモデルクラスタリングを得た．

C_1= {antelope, deer, koala, rabbit}

C_2= {buffalo, mouse, rhinoceros, tiger, wild-boar, wolf}

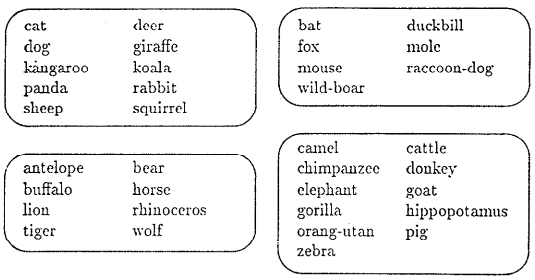


図4　因子分析による分類
Fig. 4　A classification by factor analysis.

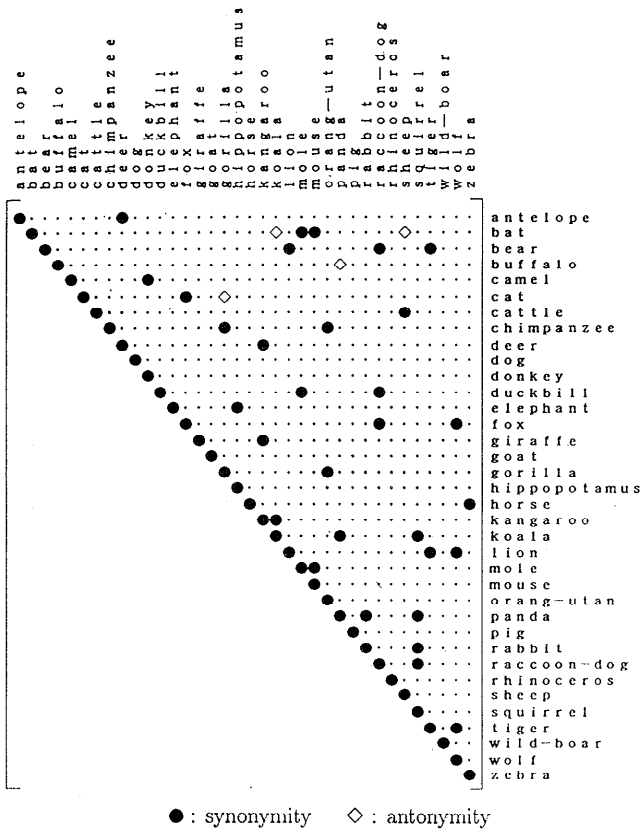


図5　類義語/反義語行列
Fig. 5　Synonymity/antonymity matrix.

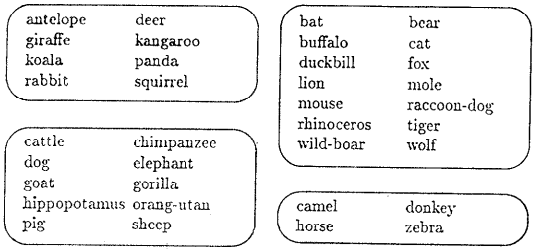


図 6 相関行列からの直観的分類
Fig. 6 An intuitive classification from a correlation matrix.

C_3= {chimpanzee, dog, goat, hippopotamus, pig, sheep}
C_4= {camel, horse}

図 6 は，本枠組によるクラスタリング結果である．因子分析による結果と比較すると，似てはいるが，微妙に違うことがわかる．

本枠組では，与えられるモデルクラスタリングの変化にともなって，生成されるクラスタリングが変化する．しかし，クラスタの数を固定するならば，事例間の部分的関係という制約のために，再編が不要になるようなモデルクラスタリングは，そういくつも存在しない．したがって，生成されたクラスタリングは，あまり意味のない多くの可能なクラスタリングのうちのひとつではない．

本枠組は，因子分析と同じように，解析結果の正当性を保証するために，何らかの実証を必要とする．すなわち，仮説を検証するためにではなくて，仮説を探索するために利用すべきデータ解析の手法である．変量間の相関係数が計算できる場合に，本枠組によって，因子分析の結果とは異なるクラスタリングが得られることは，仮説探索という観点から有意義である．

次に，本枠組を現実的なデータ解析に適用する場合の留意点について述べる．本枠組では，事例間の部分的関係を事前に準備しておく必要があるが，直観がはたらきにくい対象領域では，それが容易でない．解決方法として，たとえば，本実験で用いたような，部分的関係を相関係数に基づいて定める方法が考えられる．しかし，本枠組自体が，より適切な部分的関係やモデルクラスタリングの漸進的な獲得を支援するといえる．なぜならば，最適クラスタの定義の拡張やモデルクラスタリングの変形は，与えられるデータが，特に，解析を始めたばかりの段階では，暫定的なものであること，そして，試行錯誤を重ねることによって，次第に，妥当な結果が導き出されるようになることを前提としているからである．なお，このような試行錯誤を通じて，明確な分類の基準を発見できることがある．その場合には，必要に応じて，直観的分類と属性に基づく分類を相補的に用いることができる．

6. おわりに

本論文では，事例の直観的な分類を支援するための枠組として，事例間の類義関係と反義関係，および，モデルクラスタリングにしたがって，事例を分類する枠組を提案した．本枠組では，事例とモデルクラスタの間の親和性に基づいて，暫定的なクラスタリングを生成し，これを再編することによって，与えられたモデルクラスタリングに準拠したクラスタリング結果を導く．また，因子分析との関連について考察し，本枠組が，仮説探索のためのデータ解析の手法として有効であることを確認した．

今後の研究として，本枠組を発想支援システム[8),9)]に応用することを考えている．事例間の部分的関係から一貫性のあるクラスタリングを導く本枠組は，多変量解析に準ずるような，汎用性のあるデータ解析の手法であるが，複数の主体によって提示されたアイデアをグルーピングするという，KJ 法[5)]における作業に通じるものがある．KJ 法におけるアイデアのグルーピングは，すべて人手によるが，本枠組は，計算機による支援を可能にする．

発想支援システムでは，事例間の非類似度に基づいて，事例を空間的に配置する機能[14),15)]があると便利である．このようなマッピングの手法として，数量化理論 IV 類[3)]や多次元尺度法[4)]などの多変量解析の手法がある．しかし，KJ 法のように発想の主体が複数である場合には，行列計算などの数値的な操作によって事例の布置図を生成するだけでは不十分で，入力と出力の間の関係が，より具体的に説明されることが望ましい．その点で，クラスタリングの再編の過程が，初期のモデルクラスタリングから最終のモデルクラスタリングに至る一連の記号的な変形操作として示される本枠組においては，複数の主体による議論が容易になると考えられる．

謝辞 本研究は，産業科学技術研究開発制度「新ソフトウェア構造化モデルの研究開発」の一環として情報処理振興事業協会（IPA）が新エネルギー・産業技術総合開発機構から委託をうけて実施したものである．また，有益な助言をいただいた査読者の方々に感謝する．

参考文献

1) Aha, D. W., Kibler, D. and Albert, M. K.: Instance-Based Learning Algorithms, *Machine Learning*, Vol. 6, pp. 37-66 (1991).
2) Cheeseman, P., Self, M., Kelly, J., Stutz, J., Taylor, W. and Freeman, D.: Bayesian Classification, *Proc. 7th National Conference on Artificial Intelligence*, pp. 607-611 (1988).
3) 林知己夫：数量化の方法，p. 259，東洋経済新報社 (1974).
4) 林知己夫，鮑戸　弘編：多次元尺度解析法，p. 290，サイエンス社 (1976).
5) 川喜田二郎：発想法，p. 220，中央公論社 (1967).
6) Michalski, R. S.: A Theory and Methodology of Inductive Learning, in Michalski, R. S., Carbonell, J. G. and Mitchell, T. M. (eds.), *Machine Learning: An Artificial Intelligence Approach*, Chapter 4, Tioga (1983).
7) Mitchell, T. M.: Version Spaces: A Candidate Elimination Approach to Rule Learning, *Proc. 5th International Joint Conference on Artificial Intelligence*, pp. 305-310 (1977).
8) Munemori, J. and Nagasawa, Y.: Development and Trial of Groupware for Organizational Design and Management: Distributed and Cooperative KJ Method Support System, *Information and Software Technology*, Vol. 33, No. 4, pp. 259-264 (1991).
9) Ohiwa, H., Kawai, K. and Koyama, M.: Idea Processor and the KJ Method, *Journal of Information Processing*, Vol. 13, No. 1, pp. 44-48 (1990).
10) Quinlan, J. R.: Induction of Decision Trees, *Machine Learning*, Vol. 1, pp. 81-106 (1986).
11) Saito, Y., Tojo, S. and Komiya, S.: Intuitive Classification Based on Affinity, *Proc. 10th European Conference on Artificial Intelligence*, pp. 468-470 (1992).
12) 芝　祐順：因子分析法，p. 298，東京大学出版会 (1979).
13) Stepp, R. E., III, and Michalski, R. S.: Conceptual Clustering: Inventing Goal-Oriented Classifications of Structured Objects, in Michalski, R. S., Carbonell, J. G. and Mitchell, T. M. (eds.), *Machine Learning: An Artificial Intelligence Approach* (Vol. 2), Chapter 17, Morgan Kaufmann (1986).
14) Sumi, Y., Hori, K. and Ohsuga, S.: Computer-Aided Thinking Based on Mapping Text-Objects into Metric Spaces, *Proc. 2nd Pacific Rim International Conference on Artificial Intelligence*, pp. 183-189 (1992).
15) 田村　淳：記号間の力学に基づく概念マップ生成システム SPRINGS，情報処理学会論文誌，Vol. 33, No. 4, pp. 465-470 (1992).

付　録

A. 再編1における計算の詳細

[縮小化が可能でない理由]

(1)　$R=\{x \mid$ 二四. $Syn \cap x.Ant=\phi, x \in C_1\}$

とすると，

$C_1=$ {森鷗}

二四. $Syn \cap$ 森鷗. $Ant=\phi$

であるから，

$R=C_1=$ {森鷗}

よって，aff(二四, R) $\neq aff$(二四, C_1) $+1$ である．

(2)　$R=\{x \mid$ 二四. $Ant \cap x.Syn=\phi, x \in C_1\}$

とすると，

$C_1=$ {森鷗}

二四. $Ant \cap$ 森鷗. $Syn=\phi$

であるから，

$R=C_1=$ {森鷗}

よって，aff(二四, R) $\neq aff$(二四, C_1) $+1$ である．

(3)　$R=\{x \mid$ 二四. $Syn \cap x.Ant=\phi, x \in C_2\}$

とすると，

$C_2=$ {夏漱，芥竜，三由}

二四. $Syn \cap$ 夏漱. $Ant=\phi$

二四. $Syn \cap$ 芥竜. $Ant=\phi$

二四. $Syn \cap$ 三由. $Ant=\phi$

であるから，

$R=C_2=$ {夏漱，芥竜，三由}

よって，aff(二四, R) $\neq aff$(二四, C_2) $+1$ である．

(4)　$R=\{x \mid$ 二四. $Ant \cap x.Syn=\phi, x \in C_2\}$

とすると，

$C_2=$ {夏漱，芥竜，三由}

二四. $Ant \cap$ 夏漱. $Syn=\phi$

二四. $Ant \cap$ 芥竜. $Syn=\phi$

二四. $Ant \cap$ 三由. $Syn=\phi$

であるから，

$R=C_2=$ {夏漱，芥竜，三由}

よって，aff(二四, R) $\neq aff$(二四, C_2) $+1$ である．

以上のことから，再編1での縮小化は可能でない．

[C_1 と C_2 の統合化が可能である理由]

$U=C_1 \cup C_2$ とすると，

$U=$ {森鷗，夏漱，芥竜，三由}

aff(二四, U) $=1$

ここで，aff(二四, C_1) $=aff$(二四, C_2) $=1$ より，

aff(二四, U) $=aff$(二四, C_1) $=aff$(二四, C_2)

また，

$\forall x : (x \in Gen(C_1)) \rightarrow (aff(x, U) \geq aff(x, C_1))$

なぜならば，

$Gen(C_1)$＝{森鷗}

aff(森鷗, U)＝1, aff(森鷗, C_1)＝1

さらに，

$\forall x:(x \in Gen(C_2)) \rightarrow (aff(x, U) \geq aff(x, C_2))$

なぜならば，

$Gen(C_2)$＝{菊寛，芥竜，三由}

aff(菊寛, U)＝1, aff(菊寛, C_2)＝1

aff(芥竜, U)＝2, aff(芥竜, C_2)＝2

aff(三由, U)＝2, aff(三由, C_2)＝2

したがって，C_1 と C_2 の統合化が可能である．

B．再編2における計算の詳細

[C_4 の縮小化が可能である理由]

$R=\{x \mid$泉鏡$.Syn \cap x.Ant=\phi, x \in C_4\}$ とすると，

C_4＝{永荷，谷潤，横利，太治}

泉鏡$.Syn \cap$永荷$.Ant=\phi$

泉鏡$.Syn \cap$谷潤$.Ant=\phi$

泉鏡$.Syn \cap$横利$.Ant \neq \phi$

泉鏡$.Syn \cap$太治$.Ant=\phi$

であるから，

R＝{永荷，谷潤，太治}

aff(泉鏡, R)＝2

ここで，aff(泉鏡, C_4)＝1 より，

aff(泉鏡, R)＝aff(泉鏡, C_4)＋1

また，

$\forall x:(x \in Gen(C_4)) \rightarrow (aff(x, R) \geq aff(x, C_4))$

なぜならば，

$Gen(C_4)$＝{梶基，坂安，井靖，太治}

aff(梶基, R)＝1, aff(梶基, C_4)＝1

aff(坂安, R)＝1, aff(坂安, C_4)＝1

aff(井靖, R)＝1, aff(井靖, C_4)＝1

aff(太治, R)＝2, aff(太治, C_4)＝2

したがって，C_4 の縮小化が可能である．

C．再編3における計算の詳細

[縮小化が可能でない理由]

(1) $R=\{x \mid$永荷$.Syn \cap x.Ant=\phi, x \in C_4\}$

とすると，

C_4＝{永荷，谷潤，太治}

永荷$.Syn \cap$永荷$.Ant=\phi$

永荷$.Syn \cap$谷潤$.Ant=\phi$

永荷$.Syn \cap$太治$.Ant=\phi$

であるから，

$R=C_4$＝{永荷，谷潤，太治}

よって，aff(永荷, R)≠aff(永荷, C_4)＋1 である．

(2) $R=\{x \mid$永荷$.Ant \cap x.Syn=\phi, x \in C_4\}$

とすると，

C_4＝{永荷，谷潤，太治}

永荷$.Ant \cap$永荷$.Syn=\phi$

永荷$.Ant \cap$谷潤$.Syn=\phi$

永荷$.Ant \cap$太治$.Syn=\phi$

であるから，

$R=C_4$＝{永荷，谷潤，太治}

よって，aff(永荷, R)≠aff(永荷, C_4)＋1 である．

(3) $R=\{x \mid$永荷$.Syn \cap x.Ant=\phi, x \in C_7\}$

とすると，

C_7＝{森鷗，夏漱，芥竜，三由}

永荷$.Syn \cap$森鷗$.Ant=\phi$

永荷$.Syn \cap$夏漱$.Ant=\phi$

永荷$.Syn \cap$芥竜$.Ant=\phi$

永荷$.Syn \cap$三由$.Ant=\phi$

であるから，

$R=C_7$＝{森鷗，夏漱，芥竜，三由}

よって，aff(永荷, R)≠aff(永荷, C_7)＋1 である．

(4) $R=\{x \mid$永荷$.Ant \cap x.Syn=\phi, x \in C_7\}$

とすると，

C_7＝{森鷗，夏漱，芥竜，三由}

永荷$.Ant \cap$森鷗$.Syn=\phi$

永荷$.Ant \cap$夏漱$.Syn=\phi$

永荷$.Ant \cap$竜芥$.Syn=\phi$

永荷$.Ant \cap$三由$.Syn=\phi$

であるから，

$R=C_7$＝{森鷗，夏漱，芥竜，三由}

よって，aff(永荷, R)≠aff(永荷, C_7)＋1 である．

以上のことから，再編3での縮小化は可能でない．

[統合化が可能でない理由]

$U=C_4 \cup C_7$ とすると，

U＝{森鷗，夏漱，永荷，谷潤，芥竜，太治，三由}

aff(永荷, U)＝2

ここで，aff(永荷, C_4)＝aff(永荷, C_7)＝2 より，

aff(永荷, U)＝aff(永荷, C_4)＝aff(永荷, C_7)

ところが，

$\sim \forall x:(x \in Gen(C_4)) \rightarrow (aff(x, U) \geq aff(x, C_4))$

なぜならば，

泉鏡$\in Gen(C_4)$

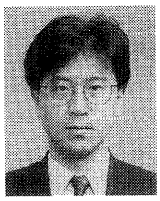
aff(泉鏡, U)＝1, aff(泉鏡, C_4)＝2

したがって，再編3での統合化は可能でない．

(平成4年11月24日受付)
(平成5年7月8日採録)

斉藤　康彦（正会員）

1961 年生．1984 年早稲田大学教育学部卒業．同年，(株)協栄計算センター（現(株)アイネス）入社．1991 年より情報処理振興事業協会（IPA）に出向．新ソフトウェア構造化モデル研究本部に所属．要求分析法の研究に従事．人工知能学会会員．



東条　敏（正会員）

1983 年東京大学工学系大学院修了．同年三菱総合研究所入社．現在情報科学部人工知能室長，主任研究員．1986〜1988 年，カーネギー・メロン大学機械翻訳センター客員研究員．自然言語理解，状況推論の研究に従事，その他人工知能一般に興味を持つ．人工知能学会，ソフトウェア科学会，ACL 各会員．

古宮　誠一（正会員）

1969 年埼玉大学理工学部数学科卒業．1970 年 4 月(株)日立製作所入社．1984 年 12 月情報処理振興事業協会（略称：IPA）に出向．以来，CASE ツールの構築技術（自動プログラミング，ソフトウェア再利用技術，プロトタイピング支援技術，分散協調開発支援技術など），仕様化／設計の方法論，ソフトウェアプロセス，CAI および知的 CAI などの研究に従事．現在は IPA の新ソフトウェア構造化モデル研究本部本部長付と技術センター特別研究員を兼務．また，1993 年 4 月より徳島大学客員教授．訳書に「ソフトウェア技術者のためのプロジェクト管理の成功への秘訣」（共訳，共立出版）がある．知能ソフトウェア工学研究会（電子情報通信学会）監事．