3F1-04

状況分解が抽出する概念の性質についての考察

What is the nature of the concept which is extracted from situation decomposition?

山川 宏[*1]
Hiroshi Yamakawa

[*1](株) 富士通研究所
FUJITSU LABORATORIES LTD.

Situation decompositions are data analysis methods, which extract plural situation from standard form data. Each extracted situations are pairs of case-subset and variable-subset, which has regularity in them. Beginning idea of this research is "How to extract partial information which constitutes real world." Since processing of situation decomposition extracts the local maximum of the evaluation function for each situation, the evaluation function determines most nature of the extracting situations. The present criteria were designed to solve a specific problem, so they are not enough generality. For the purpose of future evaluation function designing, nature of a situation is clarified, in this report. First, many factors are recognized through comparison of Matchability and ETMIC criteria. Next, a relation with other principles to which many factors relate is considered. At the last, the possibility that the frame concepts for ontology can be selected using situation evaluation function, is suggested.

1. はじめに

概念学習の分野には「例からの学習」と「観察による学習」があり，状況分解を含む，観察による学習では，システムに与えられる事例が，どの概念に属するのかがわかっておらず，さらには概念の存在すらもわかっていないのが特徴である [半田 92].

我々が提案している，状況分解は，図 1 の概念図に示すように，定型データ (変数の集合と事例の集合の直積集合) から，部分変数と部分事例のペアである状況という概念を複数個抽出する [山川 99]．状況分解は，部分事例と部分変数を選択する点で概念形成 (概念クラスタリング) と類似している．しかし，事例選択の目的が分類でないため,「状況の事例選択が相互に排他的でない点」,「状況の評価にデータ全体から寄与が含まれない点」，などが概念形成と異なる．

一方で，パターン認識の前処理等に用いられる，観察による学習に類する特徴量選択では，互いに独立性の高い変数を選ぶ傾向をもつのに対して，逆に状況分解では，状況ごと関係が強くなる (独立性の低い) ように，変数の部分集合を抽出する．

状況分解が抽出する状況の性質は, それを評価する状況分解基準によりほぼ決定される．しかし，これまでの基準の設計は特定の課題を解決するよう構成的アプローチを採ったため一般性が不十分であった．

本稿では，状況分解基準の一般性を高める準備として，その性質の明確化を行う．3 章以降で，既存の状況分解基準である Matchability 基準と ETMIC 基準の比較検討から，基本的な諸要因を整理し，似た性質を持つ諸原理との関連づけを行う．

2. 状況分解

2.1 状況分解の基本的なアイディア

状況分解という研究の，出発点となったアイディアは,「実世界情報は関係をもつ部分情報の集合で構成されるので，それを分解して取り出せれば，複数視点からの予測などに有用であろう」という，考えであった．例えば，計算機端末をマウスで操作する場合に，ポインタの位置とマウスの動きはマウスを手で掴んでいるときには強い関係を持つが，手を離した場合にはその関係はなくなってしまう．

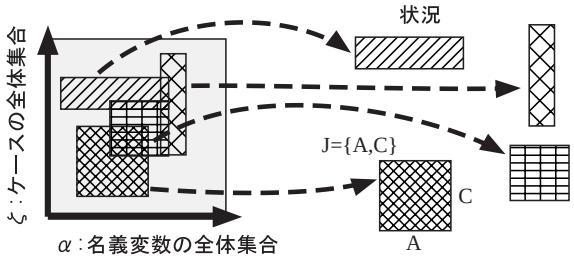

図 1: 状況分解の概念図
事例／変数それぞれの部分集合の選択を表現しているが図での表示が困難なため，便宜的に連続領域で示す.

図 1 のように，取り扱う情報を，事例の集合である定型データと考えると，全事例に同じような関係が存在すると考えるのではなく，部分的な事例において部分的な変数間に強い関係が存在しているとみなす．つまり，状況分解の目的は，一見複雑な情報を、比較的単純な関係を内在する複数の情報源に分解するために，部分事例と部分変数のペアである “状況” を抽出することにある．しかし，個々の情報源に対する詳細な概念の特定は行わない．

状況分解の振る舞いを説明するために，図 2 にデータ分布の例を示す．3 次元変数空間中の各変数 X, Y, Z の変域を $[0.0, 1.0]$ とし，事例は 1,000 個である．平面 A($X + Z = 1$) 上に 500 事例が，平面 B($Y + Z = 1$) 上に 500 事例が，それぞれ一様に分布する．状況分解は各変数を離散化して名義変数として扱うため，分布に関する線形性は考慮されない．

状況分解は，内部に関係を持つ部分状況を抽出しようとするため，図 2 に示すように平面 A 上の事例と変数 X, Z 及び，平面 B 上の事例と変数 Y, Z の選択が望まれる．また，2 平面の交線上の事例と変数 X, Y, Z が選択されてもよい．図 1 で言えば，平面 A の抽出は名義変数の全体集合 α からの部分変数 X, Z の選択と，事例の全体集合 ζ からの平面 A 上の 500 事例の選択である．

連絡先: 山川宏，(株) 富士通研究所 IT コア研究所,
〒 211-8588 川崎市中原区上小田中 4-1-1，tel:044-754-2658, fax:044-754-2693, e-mail:*ymkw@jp.fujitsu.com*

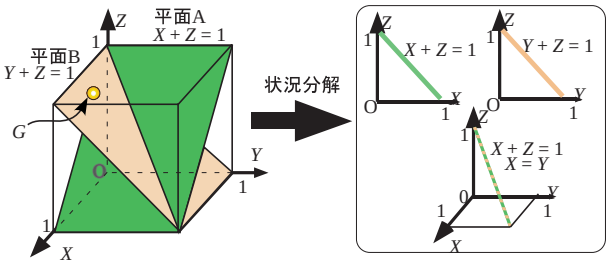

図 2: 状況分解の例題

状況分解の処理は，“状況” を評価する状況分解基準の値が，極大となる複数の状況を選択することである．これまで状況分解基準の設計は，上記のような例題などで，所望の状況が得られるように，ボトムアップなアプローチをとっていた．

本稿では，今後の一般性の高い状況分解基準の設計を目指し，状況の性質の明確化を行う．

2.2 状況分解の定式化

2.21 **名義変数とその集合:** 名義変数の全体集合を α とし，その i 番目の名義変数を $(a_i \in \alpha : i \in \{1, 2, \ldots, |\alpha|\})$ とする．任意の名義変数の部分集合を A とし $(A \subset \alpha)$，そのサイズを $|A|$ とする．

部分変数集合 A の最近傍の変数集合を，i 番目の変数選択を変化させた $A^{\pm i}$ とする．これは未選択の i 番目の変数を追加した部分変数集合 $A \cup a^i$ と，逆に選択済みの i 番目の事例を非選択とした部分事例集合 $A - a^i$ との 2 種類である．

2.22 **部分空間:** i 番目の名義変数 a_i のみにより指定された部分空間を X_{a_i} とし，そこに含まれるある値を $x_{a_i} \in \chi_{a_i}$ とする．ここで，χ_{a_i} は部分空間 X_{a_i} のドメイン (値の全体集合) である．これを，一般化して，名義変数の部分集合 A により指定された部分空間を X_A とし，そこに含まれるある値を $x_A (\in \chi_A)$ とする．ここで，χ_A は部分空間 X_A にドメイン (値の全体集合) である．

図 3 の例で示すように，部分空間 X_A のドメイン数 $|\chi_A|$ は，選択された名義変数のドメイン数 $|\chi_{a_i}|$ の積で，次式となる．

$$|\chi_A| = \prod_{a_i \in A} |\chi_{a_i}| \tag{1}$$

このとき，部分空間 X_A に一つの名義変数を追加および削除した場合の部分空間には以下の関係がある．

[追加] $X_{A \cup a_i} = X_A \times X_{a_i}$

[削除] $X_A = X_{A - a_i} \times X_{a_i}$

2.23 **事例:** 事例は，全て ($|\alpha|$ 個) の名義変数 a_i に対する変数値のベクトルであり，事例番号 j により指定される事例を $c^j = (x^j_{a_1}, x^j_{a_2}, \ldots, x^j_{a_{|\alpha|}})$ とする．事例の全体集合を ζ とすれば，$c^j \in \zeta : j \in \{1, 2, \ldots, |\zeta|\}$ である．

事例の部分集合である部分事例集合を $C (\subset \zeta)$ とし，そのサイズを $n_C (= |C|)$ とする．

事例 c^j の部分空間 X_A への射影関数を T_{X_A} とし，部分空間 X_A に射影された j 番目の事例 c^j を $c^j_{X_A} (= T_{X_A}(c^j))$ と記述する．これはたとえば，全変数 $|\alpha|$ が 2 以上の高次の空間から図 3 に示すような 2 次元空間に射影する関数である．

部分事例集合 C の最近傍の部分事例集合を，j 番目の事例選択を変化させた $C^{\pm j}$ とする．これは未選択の j 番目の事例を追加した部分事例集合 $C \cup c^j$ と，逆に選択済みの j 番目の事例を非選択とした部分事例集合 $C - c^j$ との 2 種類である．

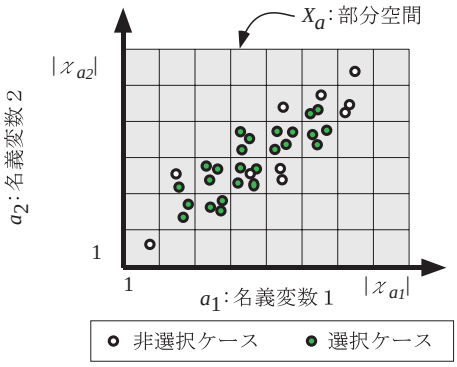

図 3: 部分空間と事例
これは 2 次元名義変数の場合の例である．

2.24 **事例分布と状況:** 部分事例集合 C が，部分空間 X_A において値 x をとる頻度分布を $F_{X_A}(x|C)$ とすれば，これは部分事例集合 C に含まれる事例の数を値毎にカウントすることで得られるので，δ をデルタ関数として，

$$F_{X_A}(x|C) = \sum_{c \in C} \delta(x = T_{X_A}(c)) \tag{2}$$

であり，この値に対応する頻度確率 $P_{X_A}(x|C)$ を

$$P_{X_A}(x|C) = \frac{F_{X_A}(x|C)}{n_C} \tag{3}$$

で得ることができる．また，ある部分空間で，ひとつ以上の事例を持つ値の数を r_A とする (図 3 参照)．

$$r_A(C) = \sum_{x \in \chi_A} \theta(F_{X_A}(x|C)) \tag{4}$$

ここで，$\theta(u)$ は,$u > 0$ で 1, $u \leq 0$ で 0 を採る階段関数．

図 1 に示すように，$J = \{A, C\}$ は部分名義変数集合 A と部分事例集合 C により指定される．この時，全体状況は全名義変数と全事例を含むので $\{\alpha, \zeta\}$ である．

状況分解の基本的な処理は，状況 J に対する評価である状況分解基準数を用いて，名義変数と事例の各々のベキ集合についての直積集合に相当する，$2^{|\alpha|+|\zeta|}$ の広さを持つ探索空間から．評価が極大となる状況を選択する．

3. 状況分解基準の性質について

状況分解基準が状況の性質を左右するために，従来のボトムアップな状況分解基準の設計には，一般性の点で問題があった．一般性の高い状況分解基準を設計する準備として，本章では，既に提案した，Matchability 基準と ETMIC 基準の比較検討から，状況という概念が持つ性質を議論する．

3.1 状況分解基準の基本 5 要因

状況分解は，関係の強い部分情報である状況を，事例と変数の両面から選択する．一つの状況 J に着目すれば，事例は，状況 J に含まれる事例 (選択事例) と， 状況 J に含まれない事例 (非選択事例) に分類される．そして，変数は，状況 J に含まれる共起/共変関係を持つ変数 (選択変数) と， 状況 J に含まれない汎化可能な変数 (非選択変数) に分類される．

事例数増大:	状況 J に含む事例数が多い.
変数数増大:	状況 J に含む変数数が多い.
共起性増大:	状況 J 内での共起関係を強くする.
共変性増大:	状況 J 内での共変関係を強くする.
変数独立性:	状況 J の非選択変数からの独立性を高める.

図 4: 状況 J に対する評価基準の基本 5 要因

次説で述べる，具体的な状況分解基準は，本質的に図 4 に記する基本 5 要因を含む．選択事例に関しては，事例数増大により，状況の再利用性を向上させる要因がある．非選択事例に関する要因は特に無い．

選択変数に関して，変数数の増加に対する変数数増大要因と，二種類の選択事例同士の関係に応じた要因がある．共起性増大は，状況 J 内での，“共起関係” を増加させる要因である．“共起関係” は，二つの出来事が同時に起こる関係で，状況 J 内で不変なので状況の特定に有用である．これは類似関係，相同関係，同時生起関係などとも呼ばれる．共変性増大は，状況 J 内に，“共変関係” を増加させる要因である．“共変関係” は，一方が変われば他方も変わる関係で，状況 J 内で変化するため，状況を利用した予測に役立つ．共変関係は，名義尺度では連関関係，順序尺度以上の変数では相関関係と呼ばれる．

そして，変数独立性は，選択事例内の関係を，非選択変数に対して汎化する妥当性が高める要因である．

3.2 具体的な基準とそこに含まれる諸要因

本節では，基本 5 要因が，具体的な 2 つの状況分解基準にどのように含まれるかを考察する．

3.21 **Matchability 基準:** Matchability 基準は，部分状況 $J = \{A, C\}$ に対する下記の評価値である [山川 99].

$$M(A,C) = M(n_C, r_A, |\chi_A|) = \mathsf{C}_1 \log \frac{n_C}{|\zeta|} + \mathsf{C}_2 \log \frac{n_C}{r_A} - \mathsf{C}_3 \log \frac{r_A}{|\chi_A|} \quad (5)$$

$\mathsf{C}_1, \mathsf{C}_2, \mathsf{C}_3$ は正の定数

Matchability 基準は，下記の 3 要因を含む．共起性増大に，選択ドメイン数 r_A(式 4 参照) を用いる．事例数増大に，選択事例数 n_C を用いる．変数数増大に，総ドメイン数 $|\chi_A|$ を用いる．ここで，いずれの名義変数においてもドメイン数 $|\chi_{a_i}|$ は 2 以上であるから，総セグメント数 $|\chi_A|$ と選択名義変数 $|A|$ は常に単調増加の関係にある．なお，変数独立性を考慮しないので，非選択変数に対する汎化能力は期待できない．

3.22 **ETMIC 基準:** ETMIC 基準は，部分状況 $J = \{A, C\}$ に対する下記の評価値である [山川 02].

$$E(A,C) = n_C \left[\min_i \left(I_{X_{A-a_i}; X_{a_i}}(C) \right) - \max_j \left(I_{X_A; X_{a_j}}(C) \right) \right] \quad (6)$$

$H_{X_A}(C) \equiv -\sum_{x \in \chi_A} P_{X_A}(x|C) \log P_{X_A}(x|C)$ は，部分空間 X_A における，部分事例集合 C のに対するエントロピー，$I_{X_A; X_B}(C)$ は二つの部分空間 X_A, X_B に対する相互情報量[*1].

ETMIC 基準は 3 要因を含む．共変性増大として角括弧内の第一項の選択変数間の相互情報量の最小値を用い，事例数増大として選択事例数 n_C を用いる．変数独立性は，角括弧内の第二項での相互情報量の最大値によりを用いる..

*1 $0 \log 0 = \lim_{t \to 0} t \log t = 0$ とする.

3.3 関連する様々な原理での基本 5 要因

本説では，構造・モデル・ルールなどの抽出に関する諸原理を，基本 5 要因の観点から関連する考察する．

3.31 **構造主義哲学:** 構造主義哲学では，“適用場面の多寡” という傾向と，対立的関係，相同的関係，同時的共変関係，継時的共変関係などの関係に基づいて構造を抽出すると考える．相同的関係は同一性や類似性に基づく関係であるのに対して，同時的共変関係は” 身長と年齢” のような関係，継時的共変関係は時間方向の関係である．[高田 97] “適用場面の多寡” により，事例数増大を，相同的関係で共起性増大を，同時的共変関係と継時的共変関係で共変性増大を考慮する．

3.32 **ルール抽出:** 連想ルール，決定ルールなどの，多くのルール抽出技術は，頻出する共起関係を抽出するので，事例数増大と共起性増大を考慮する．ある種の連想ルールでは χ 二乗検定を取り込んで共変性増大も考慮する．ルール抽出技術は裾野が広いために一概には言えないが，典型的には変数数増大や変数独立性は考慮されず，構造主義哲学に近い傾向をもつ．

3.33 **類似性と構造整列説:** 認知科学分野の類似性の研究で知られた構造整列説では，並列結合性と 1 対 1 写像の制約を満たす，二つの構造的表象間の写像を選択する．構造内の変数や，変数間の関係が一致する写像が優先的に選択され，さらに “システム性原理” により,「高次の変数間構造が優先的に写像される．また一次の関係の中でも高次の変数間構造の引数であるものが写像される.」[鈴木 96]．ここでは，2 つの対象間の共起性増大および変数数増大の要因が考慮され，変数間構造同士を比較する点では状況分解基準より多くの内容を含んでいる．しかし，2 つの対象間の比較を前提とするので，事例数増大，変数独立性，変関係性増加などは含まれない．

3.34 **因果関係:** x が原因で y が結果となる因果関係には，(1) ある変数 x と他の変数 y との間に共変関係があり，(2) 他の変数 z の影響を統制しても当該の x と y との変数間に共変関係が認められ，(3) 当該の変数間のうち片方の変数 x が時間のうえで先行して変化しもう一方の変数 y が時間の上で後から変化する条件が必要である．(1) が共変性増大に対応し，(2) が変数独立性に対応する．

3.35 **モデル選択基準:** 統計的学習では，所与のデータを説明するユニークなモデルを選択するために，AIC,MDL 等の，広くオッカムの剃刀と呼ばれるモデル選択基準を用いる．この種の基準では，与えられた情報を可能な限り忠実に説明する尤も単純なモデルを選択する．

単純なモデルは，所与のデータ内の関係を反映するので，共起性増大と共変性増大の要因が含まれる．典型的なモデル選択は，所与の固定的な変数と事例に対する，ユニークなモデルの選択である．このため，変数数増大，事例数増大は考慮しないと考えられるが，“忠実な再現” の傾向を広く解釈すれば，これら要因を含むとも考えうる．なお，部分モデルに対する未選択の変数は想定していないため，変数独立性は考慮しない．

4. 状況分解の存在論学習への展開

本章では，存在論 (オントロジー) の主要概念の一つであるフレームに，状況分解で得られる状況を対応させることを試みる．すると，フレームの選択基準として ETMIC 基準が有望であり，一種の存在論学習に関連することを示唆する．なお，本章では，前章の 3.33 でも述べた構造整列説の観点からフレームを捉える．

4.1 構造整列説を説明するフレーム

存在論は，哲学の分野では「存在に関する体系的な理論」，情報科学の分野では「概念化 (対象とする世界に存在すると考える概念とそれらの間の関係) の明示的な規約」などと捉えられる．以下では，状況分解と対応付け易い，フレームを要素に用いる存在論について述べる．フレームはいくつかのスロットと呼ばれる変数を持ち，インスタンスではスロットに値が書き込まれている．フレーム理論では，スロットに入る値は典型値や，暗黙値であり，フレーム同士は互いにポインタで指示しあうフレームネットワークを構成する．[土屋 03]．

一方，存在論はしばしば存在論木により表現される．構造整列説に立脚して類似性を調べた認知実験によれば，存在論木における二つの名詞間の階層上での距離を変化させると，共通性と整列可能な差異の数は，提示された二つのアイテムの階層上での距離が開くにつれて減少する [Markman 00]．これは，存在論上の区別が大きくなるほど，構造整列可能な変数が減少することに相当するだろう．ここで，フレームに当てはまる事例 (インスタンス) 同士は，並列結合性と１対１写像が成り立つので，構造整列可能である．そこで，存在論木の各概念をフレームとみなせば，上記の実験結果は，比較可能なスロット数の減少として捉えられる．引き続く議論では，このように構造整列の観点から捉えたフレームを扱う．

4.2 フレームと状況の対応

変数と事例の部分集合である状況は，上記のフレームとよく対応する．フレームが持ついくつかのスロットは，状況の選択変数に対応できる．また，フレームが一部のインスタンスにのみ適用可能な点は，状況の選択事例に対応する．

また，存在木の階層間では *is-a* 関係が成り立つが，状況分解で得られた状況間にも，同様の関係が成り立つ．なお，状況分解は，所与の情報から，複数の階層を抽出しうる．

4.3 状況分解基準はフレームの選択基準になり得るか

世界をモデル化するときに，フレームの候補は大量に存在するので，有用なフレームを選択するために選択基準が必要である．ここでは，フレーム選択の基準を基本 5 要因の観点から議論する．

構造整列説に基づいたフレームなので，既に前章の 3.33 で述べた変数数増大要因を含むのは明らかである．また，フレームに含まれる事例が多いほうが，再利用する価値が高まるので事例数増大要因を含むのも妥当であろう．

次に，フレーム選択における，二つの関係性要因を検討するため，フレームに含まれるべきスロットの性質を考える．スロットはインスタンス間で比較可能な変数なので，類似性判定の視点を与える．例えば，物理的実態フレームと出来事フレームの対比では，時刻スロットは，出来事フレーム内では事例の比較に有用だが，物理的実態フレーム内での事例比較には有用でない．質量スロットなどはこの逆である．また，物体フレームと物質フレームの対比を考えると，形状スロットは物体フレーム内の比較にのみ有用である．

実際には，物理的実在フレームにおける時刻スロットや，物質フレームにおける形状スロットも想定できるが，これらは，変数値が分散しており物理的実在や物質の特定には役立たない．逆にいえば，変数値が凝集しているスロットが有用で，関係性の増大がフレーム選択の要因となる．さらに詳しく見ると，フレームに適合するインスタンス同士を分類するには共変性増大要因が必要である．一方，共起性増大要因は必要ないかもしれず，更なる検討を要する．

フレームを単に記述と見なすなら変数独立性要因は不要かもしれないが，フレームを用いた予測や，制御モジュールとしての利用を想定すると，汎化能力や再利用性を得るためにこの要因も必要であろう．

以上の議論から，フレーム選択の基準としては事例数増大, 変数数増大, 共変性増加を含むことが必須で，変数独立性と共起性増加については，検討の余地がある．ETMIC 基準は，必須の 3 要因と，変数独立性を含むので，フレーム選択基準の有力な候補となる．つまり，ETMIC 基準による状況分解は，存在論学習への一つのアプローチになりえる．

5. おわりに

状況分解は，定型データから，内部に規則性がある事例と変数の組合せである状況を，複数個抽出するデータ解析技術である．そして，抽出する状況の性質は，主に状況を評価する状況分解価基準により決定される．本稿では，状況分解基準として既に提案された Matchability 基準と ETMIC 基準の比較検討から，状況分解基準は本質的に，事例数増大, 変数数増大, 共変性増大, 共起性増大, 変数独立性の基本 5 要因を考慮していることを明確化し，諸分野の原理との関連づけを行った．

存在論を構成するフレームを，構造整列説の視点から見れば，状況分解はフレーム選択の手法と捉えられる．状況分解は，変数を固定して与える “固定表象アプローチ” であるため，存在論構築のフレーム抽出という観点からは，柔軟性が不足しているが，選択すべきフレームの評価基準に対する方向性に示唆を与えることができた．

今後は，基本 5 要因に関する知見を利用し，トップダウンな状況分解基準の設計を行いたい．

また，状況の述語論理の観点からの解釈も試みたい．選択変数を項，共変関係を述語と見なせば，状況を文とみなせそうである．状況自体の活性度は，他の状況からは項として扱えるので，一種の高階述語論理を記述できる可能性がある．

参考文献

[Markman 00] Markman, A. B.: 類似から見た心, 類似性における構造整列とその影響, pp. 68–97, 認知科学の探求, 共立出版 (2000).

[高田 97] 高田明典：知った気でいるあなたのための構造主義方法論入門, 夏目書房 (1997).

[山川 99] 山川, 岡田, 渡部：規則性を持つ部分データを抽出するアルゴリズムの提案, 1999 年情報論的学習理論ワークショップ (IBIS'99), pp. 75–80 (1999).

[山川 02] 山川宏：ETMIC 基準を用いた状況分解の大腸菌たんぱく質局在サイトデータへの適用, 第 5 回情報論的学習理論ワークショップ (IBIS2002), pp. 220–225 (2002).

[土屋 03] 土屋, 中島, 中川, 橋田, 松原, 大沢, 高間 (編)：ＡＩ事典 第２版, 共立出版 (2003).

[半田 92] 半田, 松原：不均一な領域を対象とした概念形成システム CAFE, 人工知能学会誌, Vol. 7, No. 6, pp. 49–59 (1992).

[鈴木 96] 鈴木宏明：類似と思考, 認知科学モノグラフ, No. 1, 共立出版 (1996).