連続値データ群の統合的評価を文脈に応じて実現する意味的連想検索方式

池田知弘[†] 清木康[††]

本論文では，連続値を属性とする属性値によって実体の特性が評価されたデータ集合を対象として，問い合わせとして与えられる文脈に応じた意味的な相関に応じて，実体の統合的評価を行なう検索方式を提案する．社会・経済学の分野では，評価対象の状態や価値の把握を目的として，数量化された特性に関するデータを一定の価値尺度へと投影させた指標をもとにした評価手法が一般的に用いられている．このような評価手法に加え，分析者あるいは検索者の様々な視点に応じて，指標データを対象とした評価および検索を行なう機構の実現が重要である．特に，インターネットなどの広域ネットワーク上で指標データの評価検索機構を構築する場合，インタラクティブな問い合わせ評価・検索機構の実現が重要となる．提案方式では，言葉と言葉の意味の近さを文脈に応じて定量的に扱うモデルとして提案されている意味の数学モデルに基づいた意味的連想検索方式を指標データを対象とした検索へ適用することにより，多面的な視点に応じた指標データの統合的な意味的連想検索を実現する．本論文では，提案方式を地域の豊かさを評価した指標へ適用し，地域イメージの多面的評価を行なう検索機構を示す．また実際の検索実験を通して，提案方式の有効性を明らかにする．

A Semantic Associative Search Method for Continuous Values to Obtain Combinatorial Estimations according to Contexts

TOMOHIRO IKEDA[†] and YASUSHI KIYOKI[††]

In this paper, we present a new semantic associative search method for retrieving entities by a semantic evaluation mechanism for data sets which are estimated as attributes with continuous values. The important feature of this method is to provide a context-dependent semantic evaluation mechanism for those attributes. In social and economical fields, for recognizing situations and values for objects in evaluation, evaluation models are usually used. In those models, measured data are projected into standard values as indicators for the evaluation. For providing a retrieval mechanism for indicators in the wide area network as Internet, it become important to realize retrieval functions with interactive facilities. The method proposed in this paper is based on the Mathematical Model of Meaning. By applying this model to indicators, semantic associative search for indicators is realized according to multiple view points. In this paper, we show the implementation for an application of the semantic associative search method to evaluation for regional affluence. We also show several experimental results to clarify the feasibility and effectiveness of our method in this application.

1. はじめに

近年，社会・経済学分野において，対象の実態や特色を把握するための手段として，指標を用いた評価体系が一般的に用いられている。指標とは，多数の状態変数によって規定されている対象（実体：エンティティ）が持つ特性の内から，特に抽出したい性質や特徴をできるだけ少数の特性値に投影して分かり易く表現したものである[11]．この指標体系は，対象の多面的な側面を包括的に捉えるために有効であり，価値観の多様化した現代により適した評価手法として多くの分野で注目されている．

指標で捉えられた対象の総合的な把握を目的として，指標データの統合的な分析評価を行なうための手法が用いられてきた．指標データの表現においては，距離や重量などの異なる単位で示される数値は，統一された計量尺度上の数値へと変換される．すなわち，指標データの統合的な評価分析は，標準化された尺度上の数値を対象とした演算処理により行なうことができる．従来から用いられている指標データの統合的な分析手法には，一般的な統計手法であるクラスター分析や主成分分析，因子分析，重回帰分析などの多変量解析に基づいて指標デー

† 慶應義塾大学 政策・メディア研究科
Graduate School of Media and Governance, Keio University
†† 慶應義塾大学 環境情報学部
Faculty of Environmental Information, Keio University

タの集約化を行なう手法，および，指標に重み付けを与えた上で構造化を行ない，統合指標を形成する手法が挙げられる[11]．また，多次元尺度法を適用した分析手法[3)16)]や多重指標分析法[15]に基づく分析手法などの多変量解析法に関連した応用的手法が用いられている．これらは，指標データの分析を目的としていることから，個々の指標間の関係や構造を一意に定めているために，多面的な視点に応じた統合的な評価を実現していなかった．そのため，検索者や分析者の視点に基づく多面的な側面からの指標データの統合的分析手法が求められる．現在，インターネットを介して様々なデータが大量に流通している中で，検索者や分析者の視点に立った情報の取捨選択を可能とするインタラクティブな検索機構が重要となっている．すなわち，指標データの検索機構を広域ネットワーク上に構築する上では，指標間の相関を文脈に応じて動的に計量する機構が有効となる．

本論文では，検索者や分析者が与えたコンテキスト（以下，検索者が与える文脈を表す検索語の総称を“コンテキスト”とする）に応じた，指標データの動的な評価を獲得するための意味的連想検索方式を提案する．

データ間の意味的な関係を文脈に応じて動的に計算するモデルとして，意味の数学モデルが提案されている[6)8)9)]．意味の数学モデルに基づいた意味的連想検索方式をメディアデータ検索に適用する場合には，メディア情報の特徴あるいは属性を表す単語を組合わせたメタデータが用いられてきた[6)~9),17)]．本論文では，単語と連続値を組合わせた形式のメタデータを用いることによる，指標データを対象とした意味的連想検索を実現する方式を提案する．

本方式の特徴は，検索対象のメタデータにおいて，単語と連続値を組合わせた表現形式を採用している点である．従来方式では検索対象データのメタデータを単語の組合わせで構成する．そして，言葉の意味を定めるために行なう特徴による単語の記述においては，説明されるものであるか否かを示す値(0，1，−1)を用いた定義付けが行なわれていた．本方式ではそれを連続値を用いて記述させることにより，その値を対象とした演算処理を施すことができる．すなわち，特徴ごとの比重の和を，検索対象ごとに比較することが可能となる．これは直交空間上で重み付けされたベクトルを構成することによって実現している．

2. 意味的連想検索方式の概要

ここでは，意味的連想検索方式の概要を述べ，提案する方式の従来方式との相違点について述べる．

意味的連想検索方式は，データ間の意味的な同一性や差異性が，文脈や状況に応じて動的に変化するものであるという前提のもとに，検索語と検索対象データ間の意味的な近さを計算する連想検索方式である[6)8)9)]．文脈の概念を導入することにより，言葉の意味的関係を扱うファジイデータベースシステム[12)13)]と異なり、連想検索における曖昧性を排除している点が特徴である．また，直交空間における部分空間の選択を行なう演算を定義し，与えられた文脈に応じて動的に選択される部分空間上で意味的な関係を計算することから，多変量解析による空間生成を用いた情報検索方式[2)]と本質的に異なっている[8)]．

意味的連想検索方式では，言葉と言葉の意味的な相関を計算する意味の数学モデルを基本としている．そこでは，言葉として用いられる単語は特徴語群（フィーチャ）によって説明される（図1）．特徴に基づいて多次元の直交空間を形成するために，ある言葉が各特徴語で説明されるものであるか否かを(0，1，−1)で表される値で示し，言葉を行に，特徴を列とした行列を生成する．例えば図2のような形式となる．

次に，生成された直交空間上にマッピングさせる検索対象データのメタデータベクトルの設定について述べる．この意味的連想検索方式では，検索対象データのメタデータ（以下，“検索対象メタデータ”）は，単語を組合わせた形式で表現される（図3）．そして直交空間における検索対象メタデータベクトルを形成させるために，図2に示している行列と同形式の行列において，各基底に対し絶対値最大の成分（1もしくは−1，あるいは0）を選択する演算子を定義している．その結果，図4に示した形に従って，検索対象メタデータベクトルが

単語：特徴			
Word-A:	Feature-A1,	Feature-A2, ...	, Feature-AN
Word-B:	Feature-B1,	Feature-B2, ...	, Feature-BN
Word-C:	Feature-C1,	Feature-C2, ...	, Feature-CN
⋮	⋮	⋮	⋮

図1 特徴による単語の記述
Fig. 1 Description of features for words

単語：	特徴1	特徴2	特徴3	特徴4	...	特徴N
Word-A:	1	0	0	-1	...	0
Word-B:	0	1	1	0	...	0
Word-C:	0	0	-1	0	...	1
⋮	⋮	⋮	⋮	⋮		⋮

図2 特徴による単語の記述例
Fig. 2 An example of description for words used in the features

検索対象データ：単語				
DataItem-A:	Word-A1,	Word-A2,	...	, Word-An
DataItem-B:	Word-B1,	Word-B2,	...	, Word-Bn
DataItem-C:	Word-C1,	Word-C2,	...	, Word-Cn
⋮	⋮	⋮		⋮

図3 単語によるメタデータの記述
Fig. 3 Metadata description used in words

検索対象データ:	特徴1	特徴2	特徴3	特徴4	...	特徴N
Word-A1:	1	0	0	1	...	-1
Word-A2:	0	0	-1	0	...	1
Word-A3:	0	0	0	1	...	-1
DataItem-A	1	0	-1	1	...	-1

図4 従来方式における検索対象メタデータの特徴付けの例
Fig. 4 An example of features for terget metadata in conventional methods

検索対象データ: 単語1（連続値） ～ 単語n（連続値）			
DataItem-A:	Word-1 (Value-A1),	Word-2 (Value-A2), ...	, Word-n (Value-An)
DataItem-B:	Word-1 (Value-B1),	Word-2 (Value-B2), ...	, Word-n (Value-Bn)
DataItem-C:	Word-1 (Value-C1),	Word-2 (Value-C2), ...	, Word-n (Value-Cn)
⋮	⋮	⋮	⋮

図5 単語と連続値によるメタデータの記述
Fig. 5 Metadata description used in words and continuous values

検索対象データ:	特徴1	特徴2	特徴3	特徴4	...	特徴N
Word-A1:	20.45	-30.11	0.00	22.98	...	32.11
Word-A2:	40.02	-10.11	0.00	-10.77	...	-30.33
Word-A3:	30.22	-12.25	0.00	-20.76	...	35.32
DataItem-A	90.67	-52.47	0.00	-31.53	...	67.43

図6 提案方式における検索対象メタデータの特徴付けの例
Fig. 6 An example of features for target metadata in our method

求められる．

本論文の提案方式では，検索対象メタデータは，単語と連続値を組合わせた形式で記述される（図5）．単語と特徴の行列は，例えば図6のように示される．そして，検索対象メタデータベクトルを形成させるために，行列の各基底について絶対値最大となる符号別の成分総和を求める演算子を定義している．その結果として，図6に示した形に従って検索対象メタデータベクトルを求める．この詳細について，3.2節で述べる．

ここで，提案方式における連続値の設定は，次の条件のもとで行なわれる．

(1) 検索対象メタデータにおいて単語に付与される連続値の値は，全ての値が同じ価値尺度のもとに標準化されていなければならない．これはベクトル計算をする上で，不公平な重み付けが行なわれないために必要となる．

(2) 検索対象メタデータで用いられる言葉は，全ての検索対象メタデータ間で統一させなければならない．これは重み付けされる言葉とそうでない言葉の不規則な混在を防ぐために必要となる．

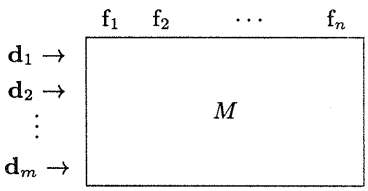


図7 データ行列 M によるメタデータの表現
Fig. 7 Metadata represented in data matrix M

3. 連続値を含むメタデータを対象とした意味的連想検索方式

ここでは，提案する連続値を含むメタデータを対象とした意味的連想検索方式を示す．

3.1 メタデータ空間 $\mathcal{MDS}$ の設定

m 個の基本データを n 個の特徴 (feature) で特徴付けることにより，特徴付ベクトル $\mathbf{d}_i(i=1,\cdots,m)$ が与えられる．そのベクトルを並べて構成した $m \times n$ 行列を M とおく（図7）．ここで，M は，列ごとに2ノルムで正規化されているものとする．このデータ行列から正規直行空間を生成し，メタデータ空間 $\mathcal{MDS}$ とする．その手順を以下に示す．

(1) データ行列 M の相関行列 M^TM を計算する．

(2) M^TM を固有値分解する．

$$M^TM = Q \begin{pmatrix} \lambda_1 & & & & \\ & \ddots & & & \\ & & \lambda_\nu & & \\ & & & 0 & \\ & & & & \ddots 0 \end{pmatrix} Q^T,$$

$0 \le \nu \le n$.

ここで行列 Q は，

$$Q = (\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n)$$

である．$\mathbf{q}_i(i=1,\cdots,n)$ は相関行列の正規化された固有ベクトル（以下，“意味素”）である．相関行列の対称性から，この固有値は全て実数であり，その固有ベクトルは互いに直交している．

(3) メタデータ空間 $\mathcal{MDS}$ を以下のように定義する．非ゼロ固有値に対応する固有ベクトル（以下，“意味素”）によって形成される正規直交空間をメタデータ空間 $\mathcal{MDS}$ と定義する．空間の次元 ν は，データ行列のランクに一致する．そしてこの空間は，ν 次元ユークリッド空間となる．

$$\mathcal{MDS} := span(\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_\nu).$$

$\{\mathbf{q}_1, \cdots, \mathbf{q}_\nu\}$ は $\mathcal{MDS}$ の正規直交基底である．

3.2 検索対象データベクトルの作成

検索対象データを表現するメタデータベクトルを形成する方法を示す．

(1) **Step-1: 検索対象データの特徴付け**

検索対象データ P に，連続値と対になったt個のメタオブジェクト（以下，“印象語”）をメタデータとして与えて，次のように定義する．α_k は連続値の値を表す．

$$P = \{\mathbf{o}_1[\alpha_1], \mathbf{o}_2[\alpha_2], \cdots, \mathbf{o}_t[\alpha_t]\}.$$

そして，各印象語 $\mathbf{o}_i$ を，メタデータ空間生成で用いたデータ行列と同一の特徴を用いて特徴付けを行なう．その際，特徴には印象語に付与されている連続値を重みとして重み付けを施す．重み付けされた特徴を，$o\_w_{i1}, o\_w_{i2}, \cdots, o\_w_{in}$ に表し，重み付印象語 $\mathbf{o}_i[\alpha_i]$ を，特徴付ベクトルとして次のように定義する．

$$\mathbf{o}_i[\alpha_i] = (o\_w_{i1}, o\_w_{i2}, \cdots, o\_w_{in}),$$
$$o\_w_{ik} = \{o_{ik}[\alpha_{ik}]\}.$$

(2) **Step-2: 検索対象データ P のベクトル表現**

検索対象データ P を構成するt個の重み付印象語 $\mathbf{o}_1[\alpha_1], \mathbf{o}_2[\alpha_2], \ldots, \mathbf{o}_t[\alpha_t]$ は，n次元のベクトルで定義される．印象語群の 和演算子 $\oplus$ を次のように定義し，検索対象データのメタデータベクトル $\mathbf{p}$ を形成する．

$$\mathbf{p} = \bigoplus_{i=1}^{t} \mathbf{o}_i := (\text{s\_sum}(o\_w_{i1}), \text{s\_sum}(o\_w_{i2}), \cdots, \text{s\_sum}(o\_w_{in})).$$

和演算子 $\bigoplus_{i=1}^{t}$ は，t 個のベクトルから各基底に対して符合別における成分の総和のうち絶対値最大の値を採用する演算子である．

$s\_sum(o_{ik})$ は，基底 k における，符合別の成分総和のうち絶対値最大の値を表す．ここで全体の総和ではなく，符合別に総和を求めることで，特徴量の絶対値数の0への近似を抑制している．

3.3 意味射影集合 Π_ν の設定

メタデータ空間 $\mathcal{MDS}$ から固有部分空間（以下，意味空間）への射影（以下，“意味射影”）の集合 Π_ν を考える．

P_{λ_i} を次のように定義する．

$P_{\lambda_i} := \lambda_i$ に対応する固有空間への射影

$i.e.\ P_{\lambda_i} : \mathcal{MDS} \to span(\mathbf{q}_i)$.

また，意味射影の集合 Π_ν を次のように定義する．

$$\Pi_\nu := \{\ 0\ , P_{\lambda_1}, P_{\lambda_2}, \cdots, P_{\lambda_\nu}, P_{\lambda_1} + P_{\lambda_2}, P_{\lambda_1} + P_{\lambda_3}, \cdots, P_{\lambda_{\nu-1}} + P_{\lambda_\nu}, \vdots \ P_{\lambda_1} + P_{\lambda_2} + \cdots + P_{\lambda_\nu}\}.$$

i次元の意味空間は $\frac{\nu(\nu-1)\cdots(\nu-i+1)}{i!}(i=1,2,\cdots,\nu)$ 個存在するので，射影の総数は，2^ν となる．つまり，このモデルは，2^ν 通りの意味の様相の表現能力をもつ．

3.4 意味解釈オペレータ $\mathbf{S}_p$ の構成

検索者の印象や検索対象データの内容を与えるコンテキストを表す ℓ 個の検索語列 $s_\ell=(\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_\ell)$ と，閾値 $\varepsilon_s(0<\varepsilon_s<1)$ が与えられたとき，それに応じた，意味射影 $P_{\varepsilon_s}(s_\ell)$ を構成するオペレータ（以下，“意味解釈オペレータ”）S_p が構成される．T_ℓ を長さ ℓ の検索語列の集合とすると，S_p は，次のように定義される．

$$S_p : T_\ell \longmapsto \Pi_\nu$$

ここで，$T_\ell \ni s_\ell, \Pi_\nu \ni P_{\varepsilon_s}(s_\ell)$.

また，$\{\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_\ell\}$ の各要素は，特徴付ベクトルであり，データ行列 M と同一の特徴で表される．オペレータ S_p は以下の計算を行なう．

(1) $\mathbf{u}_i(i=1,2,\cdots,\ell)$ をフーリエ展開する

コンテキスト s_ℓ を構成する ℓ 個の検索語を各々メタデータ空間 $\mathcal{MDS}$ へ写像する．ここで，ℓ 個の単語を各々メタデータ空間 $\mathcal{MDS}$ 内でフーリエ展開し，フーリエ係数を求める．これは，各検索語と各意味素の相関を求めることに相当する．$\mathbf{u}_i$ と $\mathbf{q}_j$ の内積 u_{ij} は次のようになる．

$$u_{ij} := (\mathbf{u}_i, \mathbf{q}_j), \quad j=1,2,\cdots,\nu.$$

ベクトル $\widehat{\mathbf{u}}_i \in \mathcal{MDS}$ を次のように定める．

$$\widehat{\mathbf{u}}_i := (u_{i1}, u_{i2}, \cdots, u_{i\nu}).$$

これは単語 $\mathbf{u}_i$ を $\mathcal{MDS}$ に写像したものである．

(2) コンテキスト s_ℓ の意味重心 $\mathbf{G}^+(s_\ell)$ を求める

まず，各意味素ごとに，フーリエ係数の総和を求める．これは，コンテキスト s_ℓ と各意味素との相関を求めることに相当する．このベクトルは，ν 個の意味素があるため，ν 次元ベクトルとなる．このベクトルを，無限大ノルムで正規化したベクトルを，以下，コンテキスト s_ℓ の意味重心 $G^+(s_\ell)$ と呼ぶ．

$$\mathbf{G}^+(s_\ell) := \frac{\left(\sum_{i=1}^{\ell} u_{i1}, \cdots, \sum_{i=1}^{\ell} u_{i\nu}\right)}{\| \left(\sum_{i=1}^{\ell} u_{i1}, \cdots, \sum_{i=1}^{\ell} u_{i\nu}\right) \|_\infty}.$$

ここで，$\|\cdot\|_\infty$ は無限大ノルムを示す．

(3) 意味射影 $P_{\varepsilon_s}(s_\ell)$ を決定する

コンテキスト s_ℓ の意味重心を構成する各要素において，閾値 ε_s を越える要素に対応する意味素を，検索対象データのメタデータを射影する意味空間の構成に用いる．意味射影 $P_{\varepsilon_s}(s_\ell)$ を次のように決定する．

$$P_{\varepsilon_s}(s_\ell) := \sum_{i\in\Lambda_{\varepsilon_s}} P_{\lambda_i} \quad \in \Pi_\nu.$$

ただし $\Lambda_{\varepsilon_s} := \{\, i \mid |\left(\mathbf{G}^+(s_\ell)\right)_i| > \varepsilon_s \}$

3.5 意味空間における相関の定量化

文脈（検索語列によって表されるコンテキスト）に対応して，3.4節で示した意味解釈オペレータ S_p を用いて選択された意味空間（部分空間）上で，その文脈と検索対象データの相関量を計算する方式を示す．

部分空間上における検索対象メタデータベクトルのノルムを求め，文脈に相関の強い検索対象データの検索を行なう．部分空間における検索対象メタデータベクトルのノルムの大きさをその文脈と検索対象データとの相関の強さとする．

コンテキスト s_ℓ が与えられた場合の検索対象メタデータベクトル $\mathbf{x}$ のノルム $\rho(\mathbf{x};s_\ell)$ を次のように定める．

$$\rho(\mathbf{x};s_\ell) = \frac{\sqrt{\sum_{j\in\Lambda_{\varepsilon_s}\cap\mathcal{S}} \{c_j(s_\ell)x_j\}^2}}{\|\mathbf{x}\|_2},$$

$$\mathcal{S} = \{i | sign(c_i(s_\ell)) = sign(x_i)\},$$

$$c_j(s_\ell) := \frac{\sum_{i=1}^{\ell} u_{ij}}{\| \left(\sum_{i=1}^{\ell} u_{i1}, \cdots, \sum_{i=1}^{\ell} u_{i\nu}\right) \|_\infty},$$

$j \in \Lambda_{\varepsilon_s}$.

ここで，意味空間を構成する意味素（固有ベクトル）群において，コンテキストに関係しているのは，正と負のどちらか一方である．そこで，意味空間を構成する意味素の符合を考慮するため，意味空間を構成する意味素の符合と正負が逆の成分についてはノルムの計算において無視している．

また，検索対象データを定義付ける特徴に与えられる値の大きさが全体的に検索対象ごとに大きく異なる場合，あるコンテキストとの相関が強いと考えられる対象ベクトルのノルムの大きさが正しく評価されず，適切な抽出が行なわれないことがある．そのために，メタデータ空間での検索対象メタデータベクトルを2ノルムで正規化することが有効となる．2ノルムで正規化をすることにより，全ての検索対象メタデータベクトルの大きさが均一化される．つまり，成分の総和である特徴量が正規化される．

ある特徴に関する検索対象メタデータベクトルの，絶対量としての大きさで検索対象を比較したい場合は，2ノルム正規化を行なわないことが望ましい．その場合、先に示した，検索対象データ $\mathbf{x}$ のノルム $\rho(\mathbf{x};s_\ell)$ は，次のように定められる．

$$\rho(\mathbf{x};s_\ell)=\frac{\sqrt{\sum_{j\in\Lambda_{\varepsilon_s}\cap S}\{c_j(s_\ell)x_j\}^2}}{\|\mathbf{x}\|}.$$

このように本方式では，検索対象のデータベクトルを2ノルムで正規化する場合と行なわない場合の二通りのメトリックを有する．

また，本方式では，意味空間における相関量の計算で，ここに示したノルムの大きさを求める方式に加えて，内積計算に基づく方式を用いることも可能である．この方式では，3.4節で示した意味重心のベクトルと，検索対象のメタデータベクトルとの内積距離を計算する．コンテキスト s_ℓ が与えられたとき，内積距離 $\bar{\eta}_{\pm}(\mathbf{x};s_\ell)$ を次のように定める．

$$\bar{\eta}_{\pm}(\mathbf{x};s_\ell)=\frac{\sum_{j\in\Lambda_{\varepsilon_s}}c_j(s_\ell)\cdot x_j}{\|\mathbf{x}\|_2}.$$

このように，ノルムの大きさおよび，内積距離を求めることにより，文脈に応じた検索対象データの相関を動的に計算する．

4. 提案方式の指標データへの適用例

ここでは，提案方式の実指標データへの適用について述べる．そして文脈に応じた意味的な相関に基づいた指標データ群の統合的評価の実現を示す．

4.1 適用させる指標データ

経済企画庁国民生活局によって編集された新国民生活指標（ＰＬＩ：People's Life Indicators）[18]を適用させる指標データとして参照する．ＰＬＩでは，国民生活における地域別の“ゆたかさ”を多面的に捉えるために，47都道府県を計144の指標で評価している．144の指標は，（住む，費やす，働く，育てる，癒す，遊ぶ，学ぶ，交わる）という8つのカテゴリーに沿って，国民生活の豊かさを網羅的に把握できるように設定されている．

ＰＬＩにおける指標データの作成では，まず，様々な調査報告から引用したデータを集計単位に考慮した上で，既に設定してある算定式に代入し、各項目に対応する数値を求めている．そして47都道府県の平均値を50とした偏差値への変換により標準化を行なっている．ここではこの値を指標データの値とする．なお，各項目に対応する数値において，その指標のレベルの上昇が個人によって望ましくないと評価される場合には，100から偏差値を減算した値を標準化指数としている．

4.2 適用の概要

指標データへの適用における，メタデータ空間の設定，検索対象メタデータの設定，文脈語列メタデータの設定について述べる．

4.2.1 メタデータ空間の設定

意味的連想検索方式におけるメタデータ空間は，検索の対象とする分野を意味的に網羅している形で生成される必要がある。例えば対象としている分野が「福祉」や「医療」などの特定分野となる場合には，固有の専門用語を特徴とさせるために専門用語集を用いることによる，特定分野に特化したメタデータ空間の生成が行なわれることが望ましい．特定分野のためのメタデータ空間生成方式については，文献10)に述べられている．

ここで適用例では，社会的な地域の豊かさを対象分野としているが，このような分野を対象としたメタデータ空間を新たに生成させることは困難である．そのため，言葉の意味的な網羅性を確保するものとして英英辞典の使用が有効となる．従って今回の適用では，これまでの意味的連想検索方式の適用実験で参照された英英辞典“Longman Dictionary of Contemporary English[19]”(以下，“LD”)を参照した．

LDでは約2000語の基本語だけを用いて約56,000語の説明がされている．このLDの基本語に合成語を加え，冠詞，be動詞，代名詞，間投詞，接続詞，前置詞，助動詞を取り除いた単語群（2,148語）(以下，“特徴語群”)をデータ行列 M の列，すなわち，特徴とした．また同じ単語群に，検索対象データに振られる言葉を加えたものを行として，2226行2148列の行列を生成した．辞典の内容をもとに，その単語を説明する特徴語が肯定の意味に用いられていた場合に“1”，否定の場合“-1”，使用されていない場合“0”とし，見出し語自身が特徴である場合その特徴の要素を“1”として自動生成した．

そして，列ごとに2ノルムで正規化を行ない，3.1節における固有値分解の際の固有値の数，すなわち意味空間の次元数は2209次元となった．

4.2.2 検索対象メタデータの設定

検索対象メタデータの設定では，3.2節で示したようにメタデータベクトルを形成させるために次の2つの条件を満たしていなければならない。

(1) 検索対象データは，連続値と対になった印象語のセットによって定義される．

(2) 各印象語は，連続値が付与された重み付の特徴のセットで特徴付けられる．

ここでは47都道府県を検索対象データとする．そして，各都道府県が評価されている指標データに対し，印象語のセットをメタデータとして設定する．すなわち，144の指標データに対して，計126の印象語を設定し，指標データの値を印象語と対となる連続値の形で付与させた．そのプロセスを示したものが，図8である．ここ

	1	2	3	4
	交通事故発生件数	公害苦情受理件数	特別養護老人ホーム施設数	有料老人ホーム定員数
北海道	58.48	68.73	65.65	50.33
青森県	52.57	54.59	60.14	41.34
岩手県	67.10	64.62	57.72	-
⋮	⋮	⋮	⋮	⋮
沖縄県	76.44	59.55	87.50	-

新国民生活指標平成１０年度版[18] より

↓

(accident = 50 + (50 - 指標 1))
(complaint = 50 + (50 - 指標 2))
(pollution = 50 + (50 - 指標 2))
(nursing_home = 指標 3 * 0.5 + 指標 4 * 0.5)
※ データがない場合は 50 を代入して計算する

↓

検索対象データ：印象語 1（連続値）～ 印象語 126（連続値）

Hokkaido：	accident(41.52),	complaint(31.27),	pollution(31.27),	nursing_home(57.99)	...
Aomori：	accident(47.43),	complaint(45.41),	pollution(45.41),	nursing_home(50.74)	...
⋮			⋮		
Okinawa：	accident(23.56),	complaint(40.45),	pollution(40.45),	nursing_home(68.75)	...

(Hokkaido) 印象語（連続値）：特徴（連続値）

accident(41.52)：	pleasant(-41.52)	damage(41.52)	expect(-41.52)	...
complaint(31.27)：	satisfaction(-31.27)	happy(-31.27)	pain(31.27)	...
nursing_home(57.99)：	care(57.99)	private(57.99)	live(57.99)	...
⋮		⋮		

図 8 指標データからの検索対象メタデータの生成
Fig. 8 Generation of metadata from data represented as indicators

で連続値の設定に主観が介入しないように，平等な重み付けを前提として計算した．4.1節で示したように，指標データが内容に応じて 100 から偏差値を減算した値になっているため，その場合は印象語の意味を損なわせないために，指標データの値を 50 で反転させた（100 から減算した）値を与えた．なお，印象語の設定では，都市の地域イメージに関する先行研究である都市の地域イメージを記載する際の語彙集[4] を参考にして行なった．

4.2.3 文脈語列（問い合わせ）メタデータ設定

検索者が与える文脈語列メタデータは，3.4節で述べた意味重心を形成する特徴付ベクトルとなるために，メタデータ空間の生成時に生成したデータ行列と同一の特徴で表される必要がある．

そこで，4.2.1節の特徴語群に，4.2.2節で用いた印象語を加えた計 2226 語を文脈問い合わせの単語とした上で，それぞれを特徴語によって特徴付けを施して文脈語列メタデータとした．

4.3 地域イメージ評価の検索システムの構築

4.2節で述べたように，メタデータ空間の生成，検索対象メタデータの設定，文脈語列メタデータの設定を行ない，地域の豊かさを示す指標データを対象とした意味的連想検索システムを構築した．このシステムでは，文脈語列メタデータで設定した 2226 語の文脈語の候補から自由に選択して組合わせた文脈語列（コンテキスト）を問い合わせとして検索が行われる．そして，文脈に応じて選択された意味空間（部分空間）上に射影された検索対象メタデータベクトルのノルムの大きさをもとに，文脈と検索対象メタデータとの相関量を計算し，結果として，意味的な相関の強さに応じて都道府県で示される地域のデータがソーティングされた形で抽出される．

図 9は，コンテキストに “comfort, relief” を与えた検索によって得られた結果である．それを視覚化（ビジュアライズ）したものが図 10である．このように視覚化することにより，検索結果の空間的な分布状況が容

コンテキスト: comfort relief

results:

Fukui	0.258148
Ishikawa	0.257438
Yamanashi	0.257223
Iwate	0.256999
Tottori	0.255178
Kagawa	0.255161
Shimane	0.255088
Tokushima	0.254887
Okayama	0.254665
⋮	⋮

図9 "comfort relief" の検索結果
Fig. 9 Retrieval result in the context "comfort relief"

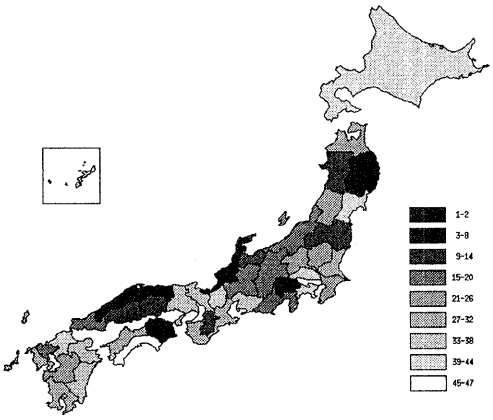


図10 検索結果のビジュアライズ
Fig. 10 Visualization of the retrieval result

易に得られるため，都道府県を包含したより広い地域レベルでの検索結果の傾向を把握することができる．このように，検索者が与えた言葉の組合せを文脈として解釈し，それに応じた意味的な相関に基づく検索が行われるため，本検索機構は地域のイメージの評価機構として捉えることができる．

5. 有効性評価実験

ここでは，4節において適用例として構築した検索機構における実験を行ない，提案方式の有効性を検証する．また，提案方式で有している複数の検索メトリックに対して検索結果の比較を行なうことにより，メトリックに応じた検索結果への影響について示す．

5.1 有効性の評価

提案方式の実現性および有効性を確認するために，指標データで示される評価対象の特性が検索結果に反映されていることを示す．そのために，指標データのメタデータ作成に用いた印象語をコンテキストとして問い合わせた検索を行ない，得られた検索対象の順位（以下，"検索結果順位"）と，指標データ値の大きさの順位（以下，"連続値順位"）の関連性を分析した．

図11は，コンテキスト "pollution" を問い合わせとして検索を行なった場合の検索結果について，連続値順位を横軸に，検索結果順位を縦軸に設定してグラフ表現したものである．順位というデータ形式には連続的な性質はないが，視覚的にわかりやすくするためにここでは折れ線グラフの形式で表現した．なおグラフ内に記されている直線は，検索結果順位が連続値順位に完全一致した状態を示しており，双方の順位の関連性が高いほどグラフとの振れが少なくなる．図11に示される結果から，双方の順位に強い相関性を確認できる．

さらに，双方の順位の関連性を定量的に評価するために，スピアマンの順位相関係数（rank correlation coefficient）を求める．スピアマンの順位相関係数（r_s）は，順位の対の数を N，順位の差を d_i として，以下のような式で表される．

$$r_s = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N^3 - N}.$$

"pollution" をコンテキストとして問い合わせた場合の，順位相関係数を計算した結果は図12に示す通りである．順位相関係数が取りうる値の範囲は0以上1以下であり，この検索では，検索結果順位と連続値順位の間にかなり強い相関を認めることができる．

ここでは，実験結果の中で比較的適した結果が得られたコンテキスト "pollution" を与えた場合について示したが，その他のコンテキストを与えた場合については5.2節に示す．

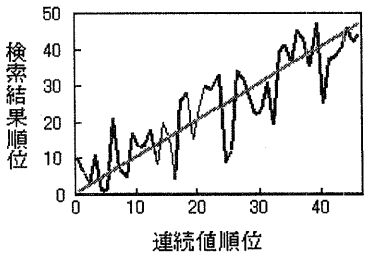


図11 "pollution" の検索結果の分析
Fig. 11 Analysis of the retrieval result in the context "pollution"

$$N = 47$$
$$\sum d_i^2 = 1258$$
$$r_s = 0.927266$$

図12 "pollution" の検索結果と順位相関係数
Fig. 12 Calculation of rank-correlation-coefficent for the result

5.2 検索メトリックの比較実験

提案方式では，より広範な検索状況に対応できるように，複数の検索メトリックを有している．それらは，3.5節で述べた検索対象メタデータの2ノルム正規化を行なう方式と行なわない方式，および，意味空間上での相関量の計算でノルム計算を行なう方式と内積計算を行なう方式である．

ここでは，各々の方式の組合わせによる4通りの計算方式における検索結果の比較実験について述べる．4.2.2節で検索対象メタデータの作成に用いた126個の印象語をそれぞれ個別にコンテキストとして問い合わせた検索を行ない，それぞれの検索結果に対して5.1節と同様にして順位相関係数を求めた．その結果を示したものが，図13である．横軸に126個の印象語，縦軸に検索結果の順位相関係数の値をとり，視覚的に捉えやすくするために折れ線形式で表している．表1は，得られた順位相関係数の平均値を示している．

これより本実験では，検索対象メタデータの2ノルム正規化を行なう方式，および，相関量計算の内積計算方式において，指標データにおける評価の値と相関性の強い検索結果が得られていることを確認できる．従って，文脈に依存した意味的な相関の強さを検索対象メタデータごとに比較して求めたい場合には，2ノルム正規化を行なうことが有効であるといえる．これは，検索対象メタデータごとでベクトルのノルムに大きな差がある時に，2ノルム正規化を行なわない場合には，どのような部分空間が選択されたとしても同じベクトルのノルムが大きくなり，適切な抽出が行なわれなくなるためである．一方，文脈に依存した意味的な相関において，ある検索対象メタデータに意味的に近い検索対象メタデータを求めたい場合には，2ノルム正規化を行なわないことが有効である．2ノルム正規化を行なうことで，検索対象メタデータのベクトルがあらかじめ有しているノルムの大きさに関する情報が除去され，検索対象メタデータごとの適切な比較が行なわれなくなるためである．

このように，アプリケーションの用途などに応じて検索メトリックの選択を行なうことが有効となることを示すことができる．

5.3 考　　察

意味的連想検索方式では，言葉と言葉の間の意味的な相関を文脈に応じて動的に計算している．提案方式の適用では，検索対象メタデータの設定として，指標データに対応した印象語を与え，さらに各印象語は特徴を用いて特徴付けが行なわれる．ここで，同じ特徴において互いに類似した特徴付けがされている印象語同士は，メタデータベクトルを形成するに当たって相互に影響が及ぶことになる．こうした作用に基づいて意味的な相関を求めているために，本意味的連想検索方式では，個々の指標データの値で示される対象の個々の特性を全く変えることなく検索結果として獲得できるわけではない．だが，検索結果の正当性が評価されるためには，指標データの値で示される対象の特性が検索結果に大きく反映さ

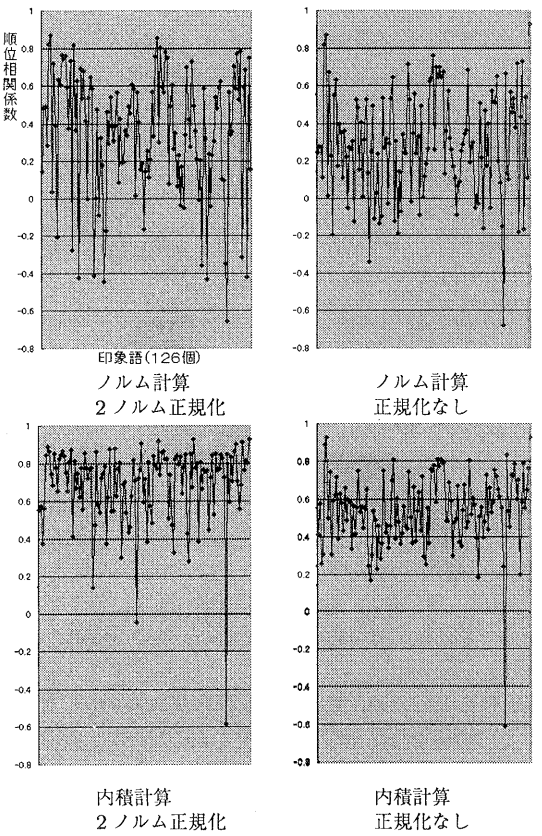


図13 順位相関係数によるメトリックの比較
Fig. 13 Comparison of the caluclation metrics by rank correlation coefficent

表1 順位相関係数の平均結果
Table 1 The average of the result as rank-correlation-coefficent

2ノルム正規化		正規化なし	
ノルム計算	内積計算	ノルム計算	内積計算
0.344963	0.690671	0.281243	0.535971

れていることを示す必要がある．

本実験では，指標データによって示される評価対象の特性と検索結果から導かれる評価対象の特性の比較を特性についての大きさの順位に基づいて行ない，両者の順位相関の強さを確認することによって，提案方式の実現性と有効性を確認することができた．

なお，評価対象が本来持っている特性をより適切に捉えた検索を行なうためには，評価対象の特性を網羅的に捉えた指標データの使用，および，より適切な印象語による検索対象メタデータの設定が有効な手段となる．

6. 指標データの統合的評価

ここでは，従来から行なわれている指標データの統合的評価の手法を示し、提案方式の適用に基づく指標データの統合的評価の特徴について考察する．

指標データの統合的な評価では，一般的な統計的分析手法として，指標データ間の相関に基づくクラスター分析や主成分分析，因子分析，および，重回帰分析などの多変量解析法を用いた指標データの集約化を行なう手法が挙げられる．

また，指標に重み付けを与えて統合指標を形成し，階層的な構造化を行なう手法がある．一般的に用いられている一対比較法（paired comparison），相対的重要度に応じた重みを直接与える手法[14]，および，指標の選好が相互に独立している上で，線形和で表される加法的な価値関数を導入することによる集約化を行なう手法[5]が挙げられる．これらの手法は，指標データ間の関係や構造を一意に定めた上で演算処理を行なう．従って，文脈に応じてメタデータ間の意味的な相関を動的に求めている提案方式とは本質的に異なっている．

さらに，多変量解析法に関連した応用的手法を挙げることができる．相関係数を類似性の測度とみなして多次元尺度法（multidimensional scaling）を適用させる手法[3][16]がある．多次元尺度法とは，複数の対象同士の類似性データをもとに，対象をできる限り少数次元の空間上に布置させる解析法である．提案方式では，文脈という概念を導入しており，直交空間における部分空間の選択を文脈に応じた意味的な相関に基づいて行なう点で，この手法と異なっている．

また，測定誤差を伴って得られた指標データから，異なる特性（抽象的概念）間の関係を探索するための手法である多重指標分析法（multiple indicators）[15]があり，この理論は文献1)における多重特性の概念に始まる．この手法は，測度の信頼性や妥当性の評価を目的としているために，提案方式が目的とする言葉の意味的な解釈に基づいた指標データの統合的評価とは主旨やアプローチが異なる．

提案方式では，指標データに対して意味という抽象的な概念に基づいて設定されたメタデータを対象に演算処理を行なう．従って，より抽象度の高いデータを定義し，操作できることから，指標データの評価を指標データの作成者から切り分けて考えることができる．これにより，指標データの作成者と評価者とが互いに独立した環境を構築することが可能となる．作成側の利点は，データの評価機能を持つ必要がなくなり，既存のデータベースや外界から取得したデータから自動的に指標データを生成する機構を考えることができることである．また，評価側の利点は，指標データをもとにして意味的な相関に基づいた情報の絞り込み検索を行なうアプリケーションを独自に構築することができることである．

7. 結　　論

本論文では，連続値データ群の統合的評価を文脈に応じた意味的な相関に基づいて獲得する意味的連想検索方式を提案した．提案方式では，直交空間を構成する各軸への重み付けに着目し，連続値データの処理機能を持つ演算子を新たに採用することにより，連続値データを含むメタデータを対象としたデータベクトルの形成を可能としている．

さらに，対象の持つ特性値を標準化された尺度に投影させる指標という評価体系に着目し，指標データを対象としたメタデータを設定することによって，提案方式の指標データへの適用を示した．適用例として，地域の豊かさを表した指標データをもとに，地域のイメージ評価検索機構を構築した．また，その機構において意味的連想検索による結果に指標データで示される評価対象の特性が反映されていることを実験により確かめ，提案方式の有効性を示した．

提案方式では，検索対象のメタデータ間の相関を検索者が与える任意の文脈に応じて動的に求めるという点で，指標データ間の関係や構造を一意に定めて分析を行なう多変量解析法とは本質的に異なっている．また，指標データの作成側にデータの評価機能を持たせる必要がなくなるため，指標データの作成者と評価者に対して，互いに独立した環境を構築することが可能となる．

今後の課題には，検索対象データのメタデータ生成に関する手法，既存の分析手法と組合わせた指標データの統合的評価手法への取り組みが挙げられる．また現在，広域ネットワーク上での指標データの編集統合を想定した，メタレベルシステムに提案方式を採り入れたマルチ・データベースシステムの設計および実現に取り組んでいる．

参考文献

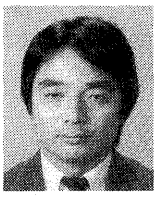
1) Campbell, D. T. and Fiske, D. W.: Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychological Bulletin*, Vol. 56, pp. 81–105 (1959).
2) Deerwester, S., Dumais, S.T., Landauer, T.K., Furnas, G. W. and Harshman, R. A.: *Indexing by latent semantic analysis*, IEEE Computer Society Press (1994).
3) Guttman, L.: Measurement as a structural theory, *Psychometrika*, Vol. 36, pp. 329–347 (1971).
4) Kasmer, J. V.: The development of a usable lexicon of environmental descriptors, *Environ. and Behav.*, Vol. 2, pp. 153–169 (1970).
5) Keeney, R. L. and Raiffa, H.: *Decision with Multiple Objectives: Preferences and Value Tradeoffs*, John Wiley & Sons, New York (1976).
6) Kitagawa, T. and Kiyoki, Y.: The mathematical model of meaning and its application to multidatabase systems, *Proc. 3rd IEEE International Workshop on Research Issues on Data Engineering: Interoperability in Mulitidatabase Systems*, pp. 130–135 (1993).
7) 清木康, 金子昌史, 北川高嗣: 意味の数学モデルによる画像データベース探索方式とその学習機構, 電子情報通信学会論文誌, Vol. J79-D-II, No. 4, pp. 509–519 (1996).
8) Kiyoki, Y., Kitagawa, T. and Hayama, T.: A metadatabase system for semantic image search by a mathematical model of meaning, *Multimedia Data Management – using metadata to integrate and apply digital media –* (Sheth, A. and Klas, W.(eds.)), McGrawHill, chapter 7 (1998).
9) Kiyoki, Y., Kitagawa, T. and Hitomi, Y.: A fundamental framework for realizing semantic interoperability in a multidatabase environment, *Journal of Intergrated Computer-Aided Engineering*, Vol. 2, No. 1, John Wiley & Sons, New York, pp. 3–20 (1995).
10) 宮川祥子, 清木康: 特定分野ドキュメントを対象とした意味的連想検索のためのメタデータ空間生成方式, 情報処理学会論文誌データベース, Vol. 40, No. SIG5(TOD2), pp. 15–28 (1999).
11) 内藤正明, 原科幸彦, 西岡秀三: 環境指標 - その考え方と作成方法 -, 日本計画行政学会 (1986).
12) Raju, K. V. S. V. N. and Majumdar, A. K.: Fuzzy Functional Dependencies and Lossless Join Decomposition of Fuzzy Relational Database Systems, *TODS*, Vol. 13, No. 2, pp. 129–166 (1988).
13) Rundensteiner, E. A., Hawkes, L. W. and Bandler, W.: On Nearness Measures in Fuzzy Relational Data Models, *International Journal of Approximate Reasoning*, Vol. 3, No. 3, pp. 267–298 (1989).
14) Stimson, D. A.: Utility measurement in public health decision making, *Manage. Sci.*, Vol. 16(2), pp. 17–30 (1969).
15) Sullivan, J. L. and Feldman, S.: *Multiple indicators*, Sage Publications (1979).
16) Weisberg, H. F. and Rusk, J. G.: Dimensions of candidate evaluation, *Amer.Political Science Review*, Vol. 64, pp. 1167–1185 (1970).
17) Yoshida, N., Kiyoki, Y. and Kitagawa, T.: An Associative Search Method Based on Symbolic Filtering and Semantic Ordering for Database Systems, *Data Mining and Reverse Engineering -Searching for Semantics-* (Spaccapietra, S. and Maryanski, F.(eds.)), Chapman & Hall, pp. 105–128 (1998).
18) 経済企画庁国民生活局: 新国民生活指標 平成10年度版, 大蔵省印刷局 (1998).
19) *Longman Dictionary of Contemporary English*, Longman (1987).

(平成11年9月20日受付)
(平成11年12月27日採録)

(担当編集委員 石川 博)

池田 知弘 (学生会員)

1976年生. 1999年慶應義塾大学環境情報学部卒業. 現在, 同大学大学院政策・メディア研究科修士課程在学中. データベースシステムの研究に興味を持つ.



清木 康 (正会員)

1978年慶應義塾大学工学部電気工学科卒業. 1983年同大学大学院工学研究科博士課程修了. 工学博士. 同年, 日本電信電話公社武蔵野電気通信研究所入所. 1984〜1995年筑波大学電子・情報工学系講師, 助教授を経て, 1996年, 慶應義塾大学環境情報学部助教授, 1998年同学部教授. データベースシステム, 知識ベースシステム, マルチメディアシステムの研究に従事. ACM, IEEE, 電子情報通信学会, 日本ソフトウェア科学会各会員.