データベースシステム 119−50
（1999. 7. 22）

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

池田知弘† 清木 康††

本稿では，連続値を属性に含むデータを対象とした意味的連想検索方式を示す．本意味的連想検索方式では，言葉と言葉の意味の近さを定量的に扱うことのできる意味の数学モデルを基礎にしており，文脈に依存した意味的な相関に基づく連想検索を可能としている．本方式の適用により，指標間の意味的な相関を文脈に応じて動的に求めることが可能となり，複数指標群の統合的な評価を多面的な視点から行なうことができる．また，評価実験の結果を示し，本方式の有効性を明らかにする．

キーワード: 意味的連想検索, 連続値, 指標データ

A Semantic Associative Search Method for Metadata containing Continuous Values

TOMOHIRO IKEDA † and YASUSHI KIYOKI††

In this paper we present a new method for performing semantic associative search for attributes with continuous values. The semantic associative search method is based on the Mathematical Model of Meanings which computes a semantic interrelation between words quantitatively, and it's possible to execute the semantic associative search by reflecting the semantic interrelation according to a given context. This method makes it possible to lead semantic interrelations between indicators according to a given context. As the result, we can obtain combinatorial estimations for several indicators from versatile viewpoints. We clarify the feasibility of the method by showing experimental results.

keywords: semantic associative search, continuous values, indicators

1. はじめに

近年，ディジタル技術の進歩や記憶媒体の普及に伴って，広域ネットワーク上に様々な形態のデータが数多く散在し，利用対象となっている．また，WWW利用の普及によって，利用者はWWWを通じて多くのメディアデータを利用可能となっている．そして現在，利用者が求めている情報を膨大なデータの中から効率的に獲得するための技術が求められている．

我々は，情報の持つ意味に着目し，言葉と言葉の意味の近さを定量的に扱う意味の数学モデルを適用させた，意味的連想検索方式を提案している[1)2)]．意味的連想検索方式では，情報を言葉の組合せで定義することにより，属性の異なる情報を意味的な相関に基づいて統合的に扱うことを可能としている．

意味的連想検索方式では，言葉の組合せで表現されたメタデータをベクトル形式で表現し，正規直交化されたメタデータ空間に写像させる．さらに，与えられた文脈に応じて選択された部分空間にそれらのベクトルを射影して距離を計算することによって，単純なパターンマッチングでは得ることのできない，文脈に依存した意味的連想に基づく情報検索を実現している．

ここで，情報の持つ意味を言葉に代表させて形式的に表現する際に，言葉の持つ比重を重み付けとして与えることで，言葉という表現形態に内包される情報の特性をより豊かに表現させることができる．

この概念に基づいて，本稿では，連続値を属性に含むデータを対象とした意味的連想検索方式を提案する．本方式では，従来の意味的連想検索方式では言葉によって表現されていた，検索対象をベクトル化するためのメタデータ形式を，言葉と連続値の対の形で表現する．言葉に連続値を付与できるようにすることで，言葉の持つ特徴量の強さを定量的な形で形式的に扱うことができる．

ここで，付与される連続値は，対となる全ての言葉に対して同じ基準を有している必要がある．意味的連想検索方式では，全ての言葉が同じ特徴量を有することを前提としているため，言葉への連続値による重みづけも平等な基準で行なわれなければならない．そのために，本方式では，対象データが属性として含む連続値を均一の

† 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

基準に基づいて標準化させるプロセスが不可欠となる．

均一の基準に基づいて標準化された連続値データの代表例として，指標データを挙げることができる．指標とは，対象の持つ特性の内から特に抽出したいものを，標準化された尺度に投影して表現したものである[5]．指標の特徴は，標準化された尺度を用いることにより，異なるデータ単位に影響されることなく，指標間の相対的な関係を考慮したデータの分析ができることである．近年，社会・経済・環境などの分野では，対象の定量的分析の手法として，指標による評価分析が一般的に行なわれるようになってきており，指標への関心が高まる一方で，その分析のあり方が強く問われている．

本稿では，提案方式に実データを適用する方法について述べ，それによって指標データの意味的な相関に基づく統合的な評価が可能となることを示す．従来における指標の統合的な評価分析では，関係式の設定[4]や，階層的指標体系の構成[5]、重回帰分析や因子分析などの多変量解析による手法が用いられている．これらの方法では，指標間の関係が一意に定められ，静的な評価が行なわれる．本稿で示す手法では，指標間の意味的な相関関係を文脈に依存した形で動的に求めるため，状況に応じた多面的な視点から評価を行なうことができる．

本稿では，実指標データを適応した実験を通じて，提案方式の有効性評価を行なう．具体的には，本方式を適用した検索機構の検索結果として，文脈に依存した意味的相関の強さに応じてソーティングされた検索対象の順位と，指標データにおける評価対象の評価値の大きさの順位とを比較する．また，その関連性を定量的に評価し，本方式の有効性を考察する．

最後に本稿では，まとめと今後の課題について述べる．本方式の特徴と本方式を適用した指標分析の特徴について考察し，今後取り組むべき課題について述べる．

2. 従来の意味的連想検索方式

本節では，従来における意味的連想検索方式の概要を述べる．詳細は文献[1)2)]に詳しく述べられている．

2.1 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1）．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．

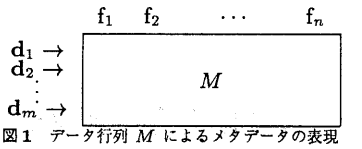

図1 データ行列 M によるメタデータの表現

(2) M^TM を固有値分解する．

$$M^TM = Q \begin{pmatrix} \lambda_1 & & & \\ & \ddots & & \\ & & \lambda_\nu & 0 \\ & & 0 & \ddots_0 \end{pmatrix} Q^T,$$

$0 \le \nu \le n$.

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．

2.2 検索対象データのデータベクトルの作成方式

検索対象データを表現するデータベクトルを形成し，メタデータ空間上にマッピングする方法を示す．

(1) **Step-1: 検索対象データの特徴付け**

検索対象データ P に，t個のメタオブジェクト（以下，“印象語”）$\mathbf{o}_1,\mathbf{o}_2,\cdots,\mathbf{o}_t$ をメタデータとして与えることにより，次のように定義する．

$$P=\{\mathbf{o}_1,\mathbf{o}_2,\cdots,\mathbf{o}_t\}.$$

さらに，各印象語 $\mathbf{o}_i$ を，メタデータ空間の生成で用いたデータ行列と同一の特徴を用いて，特徴付ベクトルとして次のように定義する．

$$\mathbf{o}_i=(o_{i1},o_{i2},\cdots,o_{in})$$

(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個の印象語 $\mathbf{o}_1,\mathbf{o}_2,\ldots,\mathbf{o}_t$ は，n次元のベクトルで定義される．和演算子 $\oplus$ を次のように定義し，検索対象データのデータベクトル $\mathbf{p}$ を形成する．

$$\mathbf{p}=\bigoplus_{i=1}^{t}\mathbf{o}_i := (\mathrm{sign}(o_{\ell_1 1}) \max_{1\le i\le t}|o_{i1}|,\ \mathrm{sign}(o_{\ell_2 2}) \max_{1\le i\le t}|o_{i2}|,\ \cdots,\mathrm{sign}(o_{\ell_n n}) \max_{1\le i\le t}|o_{in}|).$$

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絶対値最大の成分を選ぶ演算子である．sign(a) は，“a” の符号（正，負）を表す．ま

た，$l_k(k=1,\ldots,t)$ は，特徴が最大となる印象語を示す指標であり，次のように定義する．

$$\max_{1\le i\le t}|o_{ik}| = |o_{l_k k}|.$$

2.3 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．

$$P_{\lambda_i} := \lambda_i \text{ に対応する固有空間への射影}$$
$$i.e.\ P_{\lambda_i} : \mathcal{MDS} \to span(\mathbf{q}_i).$$

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．

2.4 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文脈（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，しきい値 $\varepsilon_s (0 < \varepsilon_s < 1)$ 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，無限大ノルムによって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，しきい値 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 $\Lambda_{\varepsilon_s} := \{\, i \mid \left|\left(\mathbf{G}^+(s_\ell)\right)_i\right| > \varepsilon_s \}$

2.5 意味空間における相関の定量化

メタデータ空間に写像された検索対象データ群に対応する各検索対象データベクトルについて，与えられた文脈（以下，“コンテキスト”）によりメタデータ空間から選択された意味空間（以下，“部分空間”）上で，ノルムを計算し，コンテキストと検索対象データの相関量計算を行う．意味空間における検索対象データベクトルのノルムの大きさをそのコンテキスト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データ $\mathbf{x}$ のノルム $\rho(\mathbf{x}; s_\ell)$ を次のように定める．

$$\rho(\mathbf{x}; s_\ell) = \frac{\sqrt{\sum_{j\in\Lambda_{\varepsilon_s}\cap S}\{c_j(s_\ell)x_j\}^2}}{\|\mathbf{x}\|_2},$$
$$S = \{i \mid sign(c_i(s_\ell)) = sign(x_i)\},$$
$$c_j(s_\ell) := \frac{\sum_{i=1}^{\ell} u_{ij}}{\| \left(\sum_{i=1}^{\ell} u_{i1}, \cdots, \sum_{i=1}^{\ell} u_{i\nu}\right) \|_\infty},$$
$$j \in \Lambda_{\varepsilon_s}.$$

意味空間を構成する意味素（固有ベクトル）群において，文脈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，検索対象データを特徴付ける特徴の数が検索対象ごとに異なる場合，文脈との相関が強いと考えられる対象ベクトルのノルムにおける，相対的なノルムの大きさが考慮されず，適切な抽出が行なわれないことがある．そのため，メタデータ空間での検索対象データベクトルを 2 ノルムで正規化を行なっている。

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

提案方式では，検索対象データのデータベクトルの作成方式，および，意味空間におけるコンテキストと検索対象データの相関の定量化に関するメトリックの選択を可能としている．本節では，本稿で提案する検索対象データのデータベクトルの作成方式，および，意味空間における相関の定量化について，その概要を述べる．

3.1 検索対象データベクトルの作成

(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 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に重み付けを行なった，重み付特徴 $o\_w_{i1}, o\_w_{i2}, \cdots, o\_w_{in}$ を用いて特徴付ける．重み付印象語 $\mathbf{o}_i[\alpha_i]$ を，重み付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，検索対象データのデータベクトル $\mathbf{p}'$ を形成する．

$$\mathbf{p}' = \bigoplus_{i=1}^{t}{}' \mathbf{o}_i := (\mathrm{s\_sum}(o\_w_{i1}), \mathrm{s\_sum}(o\_w_{i2}), \cdots, \mathrm{s\_sum}(o\_w_{in})).$$

和演算子 $\bigoplus'^{t}_{i=1}$ 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．

3.2 提案方式における意味空間上の相関の定量化

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．2 ノルムで正規化を行なった場合，検索対象のデータベクトルにおけるノルムの大きさが，全ての検索対象において均一化される．つまり，成分の総和である特徴量が標準化される．本方式では，それを適さないケースに考慮して，複数のメトリックを有している．

2 ノルム正規化が行なわれない場合，2.5節で示した，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．

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

また本方式では，意味空間における相関量の計算で，2.5節に示したノルムの大きさを求める方式に加え，内積計算に基づく方式を有する．この方式では，2.4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．

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

4. 提案方式の実指標データへの適用

本節では，提案方式への実データの適用方法を示し，意味的な相関に基づく指標の統合的評価を行なう．

4.1 指標データへの適用

意味的連想検索方式では，全ての言葉に同じ特徴量が与えられることを前提としており，連続値による言葉への重みづけは平等の基準で行なわれなければならない．従って，本方式では，対象データの属性に含まれる連続値を標準化させるプロセスが必要となる．

均一の基準に基づいて標準化された連続値データの代表例として，指標データが挙げられる．指標とは，対象の持つ特性の内から特に抽出したいものを，標準化された尺度に投影して表現したものである[5]．本節では，実指標データを用いて，提案方式の適用事例を示す．

4.2 メタデータ空間の設定

空間生成用メタデータの生成，すなわちデータ行列 M の生成を行なうために英英辞典，“Longman Dictionary of Contemporary English[6]”(以下，“LD”)を参照した．LD は約 2000 語の基本語だけを用いて約 56,000 語の説明をしている．

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．辞典の内容を基に，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，2.1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．

英英辞典の参照により，メタデータ生成における一貫性，客観性が保持され，意味的な網羅性を確保できる．特定分野を対象とした意味的連想検索方式を行なう場合は，固有の専門用語集を用いた空間生成が望ましい[3]．

4.3 検索対象メタデータベクトルの作成

指標データには，新国民生活指標（ＰＬＩ：People's Life Indicators）[8]を用いる．ＰＬＩは，地域の“ゆたかさ”を多面的に捉えるために，47都道府県を評価対象として，計144の指標で構成されている．

この144の指標に対して，関係が多対多になるように計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として付与した(表1)．連続値は，主観が介入しないように，平等な重みづけを前提として計算した．印象語の設定は，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[7]を参考にして行なった．

表1 標準化指数から検索対象メタデータの生成

	1	2	3
	交通事故発生件数	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	65.65	50.33
青森県	52.57	60.14	41.34
岩手県	67.10	57.72	-
：	：	：	：
沖縄県	76.44	87.50	-

新国民生活指標平成１０年度版[8]より

↓

(accident = 50 + (50 - 指標1))
(nursinghome = 指標2 * 0.5 + 指標3 * 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: メタデータ
hokkaido: accident{41.52},nursinghome{57.99} ...
  aomori: accident{47.43},nursinghome{50.74} ...
    :                    :
okinawa: accident{23.56},nursinghome{68.75} ...
```

```
印象語（連続値）: 特徴1（連続値）特徴2（連続値）...
manage{31.230}: control{31.230} charge{31.230} ...
liable{53.390}: suitable{53.390} live{53.390} ...
     :                    :
personal_computer{49.130}: computer{49.130} ...
```

4.4 文脈語列（問い合わせ）メタデータの生成

意味空間へ写像する文脈語列のメタデータを生成した．4.2節の特徴御群に，4.3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．そして，特徴語によって各々特徴付けを行なった．

4.5 地域イメージ評価の検索システム

次に，検索結果の例を示す．コンテキストに“happy, bright”を与えて検索を行なった結果，文脈に依存した意味的な相関に応じてノルムが計算され，図2に示す結果が得られた．図3は，それをビジュアライズしたものである．このように，検索者が与えた言葉の文脈に依存し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評価機構としての側面を有する．

```
lgmmm3> r on
lgmmm3> es 0.2
lgmmm3> cn happy bright :
lgmmm3> pdf2 47
results:
       okinawa     1.862751
     yamagata      1.861768
    fukushima      1.857536
       aomori      1.839770
          :            :
        hyogo      1.738801
    tokushima      1.736200
        kyoto      1.732269
```

図2 “happy bright”の検索結果

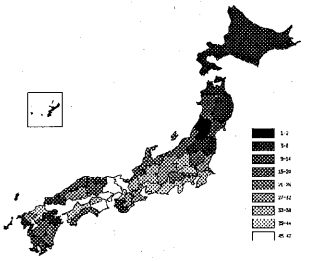

図3 検索結果のビジュアライズ

4.6 本方式にもとづく指標の統合的分析手法

提案方式への適用から，複数指標の統合的評価を実現したが，従来における手法は，主に3つに分類できる．

第一は，指標間の関係を規定した定式へ指標データを代入する，経済指標の分析で一般的に用いられる手法である[4]．この手法は時系列に沿って対象を分析する場合に有効だが，対象の詳細なモデル化には向いていない．

第二は，関連ある指標群に寄与度に応じた重みづけを行ない，階層的に統合指標を作る，社会・環境指標でよく用いられる手法である[5]．上位から下位を構築することもあり，階層レベルに応じた分析に有効だが，静的な構造ゆえに，多面的な視点に基づく分析には向かない．

第三は，重回帰分析や因子分析など多変量解析による分析手法である．様々な視点からの分析ができるが，状況の変化を動的に捉える分析には向いていない．

提案方式へ指標データを適用して実現した手法では，指標間の関係を文脈や状況に依存した動的なものとして捉える．これにより，文脈に応じて動的に変わる意味的な相関に基づいて，対象の評価を行なうことができる．

5. 有効性評価実験

5.1 有効性の評価実験

検索結果を評価するために，指標の特徴付けに用いた印象語をコンテキストに与えて検索を行ない，意味的な相関関係に応じてソーティングさ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4は，コンテキストが“pollution”の検索結果に対して，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．グラフ内の直線は，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，双方の関連性が高いほど振れが少なくなる．図4から，双方の高い相関性を確認できる．

さらに，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を求める．スピアマンの順位相関係数（r_s）は，以下のような式で求められる．

$$r_s = 1 - \frac{6\sum d_i^2}{N^3 - N}$$

N ： 順位の対の数

$\sum d_i^2$ ： 順位の差の自乗

“pollution” をコンテキストに与えた時の，順位相関係数を計算した結果を表5に示した．順位相関係数が最大に取り売る値は1であり，これより，“pollution” をコンテキストとして与えた検索では，検索結果順位と連続値順位の間にかなり高い相関を認めることができる．

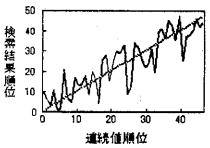

図4 “pollution” の検索結果の分析

$N = 47$

$\sum d_i^2 = 1258$

$r_s = 0.927266$

図5 “pollution” の検索結果と順位相関係数

5.2 検索メトリックの比較実験

本方式で複数有しているメトリックには，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，意味空間における相関量の計算で，ノルム計算を行なう方式と内積計算を行なう方式がある．この実験では，各々の方式の選択を組み合わせた四通りの方式に対して，4.3節で，検索対象メタデータの作成に用いた126個の印象語をコンテキストとして検索を行ない，それぞれに順位相関係数を求めた．

結果を表したものが，図6である．これらは，横軸126項目の印象語に対する，検索結果の順位相関係数の値を縦軸にとりグラフ化したものである．また，表2は，得られた順位相関係数の平均値である．この結果から，本実験では，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，指標の評価値と相関性の高い検索結果が得られていることがわかる．

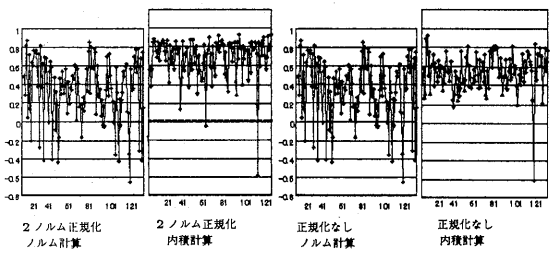

図6 単位相関係数によるメトリックの比較

表2 順位相関係数の平均

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

5.3 考　　察

5.2節の結果から，文脈に依存した意味的な相関の高さを検索対象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望ましいことが示される．検索対象ベクトルのノルムに大きな差がある場合に，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いためである．

一方，ある検索対象に文脈に依存した意味的な相関の高い検索対象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望ましい．2ノルム正規化を行なった場合には，検索対象データベクトルがはじめに有しているノルムのに大きさに関する情報が除去されるため，検索対象ごとの適切な比較が行なわれなくなるからである．

6. 結　　論

本稿では，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し，実データへの適用と性能評価を通じて，有効性の検証を行なった．本方式は，指標データへの適用により，文脈に依存した意味的な相関に応じて，指標群の統合的な評価ができる点が特徴である．

さらに，本方式では，検索対象を連続値と言葉の組合せで定義することにより，指標データに限らず，連続値を属性として有するあらゆるメディアオブジェクトを意味的連想検索の対象とすることも可能である．

今後は，対象データのメタデータ生成に関する手法，ならびに，既存の分析手法と組み合わせた指標の統合的評価手法への取り組みが課題となる．さらに，様々な属性を持つデータと連結して相互運用をはかるための，マルチデータベース環境へ本方式を組み込むことによる，広域なネットワーク上での情報編集統合を可能とする基盤整備への発展を考えている．

参 考 文 献

1) Y.Kiyoki, T.Kitagawa and T.Hayama, *“A meta-database system for semantic image search by a mathematical model of meaning,”* Multimedia Data Management – using metadata to integrate and apply digital media –,” McGrawHill(book), A. Sheth and W. Klas(editors), Chapter 7, 1998.

2) N.Yoshida, Y.Kiyoki and T.Kitagawa, “An Associative Search Method Based on Symbolic Filtering and Semantic Ordering for Database Systems,” Data Mining and Reverse Engineering -Searching for Semantics-, Chapman & Hall, S. Spaccapietra & F. Maryanski (editors), pp. 105-128, 1998.

3) 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(TOD 2),pp.15-28,1999.5.

4) 通商産業大臣官房調査統計部,「指数の作成と利用」, 通産統計協会 1994.

5) 内藤 正明, 森田 恒幸著,「環境指標の展開」, 日本計画行政学会 学陽書房 1995.

6) *“Longman Dictionary of Contemporary English,”* Longman, 1987.

7) Kasmer.J.V., *“The development of a usable lexicon of environmental descriptors, ”* Environ.and Behav.,2,pp.153-169,1970.

8) 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.