数理解析研究所講究録
915巻 1995年 75-86

意味の数学モデルとメタデータベースシステムへの応用

筑波大学　電子・情報工学系　北川　高嗣　(Takashi Kitagawa)
筑波大学　電子・情報工学系　清木　康　(Yasushi Kiyoki)
筑波大学　工学研究科　宮原　隆行　(Takayuki Miyahara)

要旨

マルチデータベース・システムにおける最も重要な課題の一つは、異なるデータベースにあるデータ間の意味の同一性、相異性の扱いである。データ間の意味的な関係の扱いについては、データ間の関係を静的かつ明示的に記述し、同一性、相異性を曖昧性を含んで判定する方法が広く用いられてきた。

我々は、単語間の意味的な同一性、相異性について、それらは、静的な関係によって決定されず、文脈や状況に応じて動的に変化するものと考える。実際には、データ間の意味的な同一性、相異性は、静的な関係によって決定されるのではなく、文脈や状況に応じて動的に変化するものであり、その動的な要素を含んで決定しなければ、データ間の関係の曖昧性を排除することはできない。このような単語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、我々は, 意味の数学モデルを提案している。この数学モデルは、マルチメディア・データベースを対象とした意味的検索（印象や直感による検索）、特に、画像検索、音楽検索に適用することができる。

本稿では、文脈あるいは状況に応じて動的に変化するデータ間の意味的な関係を計算するモデルとして、意味の数学モデルを示し、さらに、その実現方式について述べる。

We present a new method for extracting semantically related information dynamically without using explicit representations of relationships between data items. This method is used to provide a fundamental function for realizing semantical information acquisition in multidatabases systems. This method provides a function for recognizing the context and computing the equivalence and similarity between data items dynamically according to the context.

1　はじめに

マルチデータベース・システムにおいては、異なるデータベースに存在するデータ間の意味の同一性、相異性に関する扱いが重要である [1, 2, 3, 4, 12, 13, 14, 15]。現行のデータベース・システムにおける情報の抽出の基本操作は、パターン・マッチングによる検索であり、異なる表現形態であるが同一の意味をもつデータや

近い意味をもつデータによる検索を行うことはできない。また、同一のデータがもつ多義性を取り扱うことはできない。データ間の意味的な関係の扱いについては、データ間の関係を静的かつ明示的に記述し、同一性、相異性を判定する方法が広く用いられてきた。しかし、その判定は、静的に与えられた関係を用いて、曖昧性を含んで行われる。例えば、シソーラスを用いて同義語を照会する方法があるが、その同義語は、シソーラスの設計時に静的に決定され、また、同義であることの定義には曖昧性を含んでいる。すなわち、多義性のあるデータの間の意味的な関係を文脈あるいは状況に応じて動的に特定する機能を有していない。

実際には、データ間の意味的な同一性、相異性は、静的な関係によって決定されるのではなく、文脈や状況に応じて動的に変化するものであり、その動的な要素を含んで決定しなければ、データ間の関係の曖昧性を排除することはできない。

本稿では、文脈あるいは状況に応じて動的に変化するデータ間の意味的な関係を計算するモデルとして、意味の数学モデル [5, 6, 9] を示し、さらに、その実現方式について述べる。現行のデータベースシステムにおける基本操作は、パターン・マッチングによる検索を主体としている。意味の数学モデルに基づいたデータベース・システムを実現することにより、動的に変化する状況に応じた意味解釈（意味空間の選択、および、その空間内での最良近似）が可能となる。さらに、パターン・マッチングの能力を越えた、意味を考慮した柔軟な検索が可能となる。ここでは、意味の数学モデルの実現方式、および、それに従って構築した実験システムについて述べる。

2 意味の数学モデル

2.1 概要

意味の数学モデルは、言葉の意味を扱うためのモデルである。ここでは、その概要を示す。厳密な定式化については、次節において述べる。

1. **前提:** いくつかの単語を特徴づけたデータの集合が、m 行 n 列の行列 (以下、“データ行列” と呼ぶ) の形で与えられているものとする。この行列において、m 個のそれぞれの単語 (word) は、n 個の特徴 (features) によって特徴づけされている。このデータ行列の具体的な生成法については、4.2 節において述べる。

2. **イメージ空間 $\mathcal{I}$ の設定:** データ行列から、特徴づけに関する相関行列をつくる。そして、相関行列を固有値分解し、固有ベクトルを正規化する。相関行列の対称性から、この全ての固有値は実数であり、その固有ベクトルは互いに直交している。このとき、非ゼロ固有値に対応する固有ベクトル（以下、“意味素” と呼ぶ）の張る正規直交空間をイメージ空間 $\mathcal{I}$ と定義する。この空間の次元 ν は、データ行列のランクに一致する。また、この空間は、ν 次元ユークリッド空間となる。

3. **意味射影の集合 Π_ν の設定:** イメージ空間 $\mathcal{I}$ から固有（不変）部分空間（以下、“意味空間” と呼ぶ）への射影 (以下、“意味射影” と呼ぶ) の集合 Π_ν を考える。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。

4. **意味解釈オペレータ S_p の構成:** 文脈を決定する ℓ 個の単語列 (以下、“文脈語群” と呼ぶ) s_ℓ としきい値 ε_s が与えられたとする。このとき、その文脈に応じた意味射影 $P_{\varepsilon_s}(s_\ell)$ を決めるオペレータ (以下、“意味解釈オペレータ” と呼ぶ)S_p を次のように構成する。

(a) 文脈語群 s_ℓ を構成する ℓ 個の単語を各々イメージ空間 $\mathcal{I}$ へ写像する。この写像では、ℓ 個の単語を各々イメージ空間 $\mathcal{I}$ 内でフーリエ展開し、フーリエ係数を求める。これは、各単語と各意味素の相関を求めることに相当する。

(b) 各意味素ごとに、フーリエ係数の総和を求める。これは、文脈語群 s_ℓ と各意味素との相関を求めることに相当する。また、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、文脈語群 s_ℓ の意味重心と呼ぶ。

(c) このとき、文脈語群 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、単語を射影する意味空間の構成に用いる。これにより、意味射影 $P_{\varepsilon_s}(s_\ell)$ を決定する。

このオペレータは、文脈語群と相関の高い意味空間の自動的な選択を実現する。

5. **意味空間における距離計算:** 文脈語群 s_ℓ により、各意味素ごとに重みを定める。そして、意味空間において、その重みを考慮した単語間の距離計算を行う。これにより、文脈に忠実な単語間の関係の解釈が可能となる。

このモデルにより、文脈に応じた単語間の関係の解釈 (意味空間の選択、およびその空間内における最良近似) が可能となる。

2.2 具体的な定式化

本節では、前節において述べた概要の定式化について述べる。

2.2.1 イメージ空間 $\mathcal{I}$ の設定

ここでは、m 個の単語について各々 n 個の特徴
$(f_1, f_2, \cdots, f_n)$ を列挙した各単語に対する特徴付ベクトル $\mathbf{w}_i (i = 1, \cdots, m)$ が与えられているものとし、そのベクトルを並べた m 行 n 列のデータ行列を A とする。

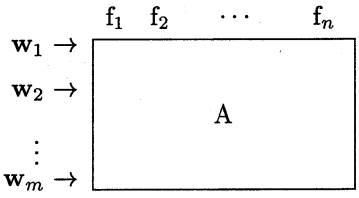


図 1: データ行列 A の構成

1. データ行列 A の相関行列 A^TA を作る。

2. A^TA を固有値分解する。

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

$$0 \leq \nu \leq n.$$

ここで行列 Q は、

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

である。この $\mathbf{q}_i$ は、相関行列の固有ベクトル、つまり意味素である。

3. このとき、イメージ空間 $\mathcal{I}$ を以下のように定義する。

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

$(\mathbf{q}_1, \cdots, \mathbf{q}_\nu)$ は $\mathcal{I}$ の正規直交基底である。

2.2.2 意味射影集合 Π_ν の設定

P_{λ_i} を次の様に定義する。

$P_{\lambda_i} \overset{\mathrm{d}}{\Longleftrightarrow} \lambda_i$ に対応する固有空間への射影、

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

意味射影の集合 Π_ν を次のように定義する。

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

Π_ν の要素の個数は 2^ν 個であり、これは 2^ν 通りの意味の様相表現ができることを示している。

2.2.3 意味解釈オペレータ $\mathbf{S}_p$ の構成

文脈語群

$$s_\ell = (\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_\ell)$$

と、正数 $\varepsilon_s (\varepsilon_s > 0)$ が与えられたとき、意味解釈オペレータ S_p は、その文脈語群 s_ℓ に応じて、意味射影 $P_{\varepsilon_s}(s_\ell)$ を決定する。すなわち、$s_\ell \in T_\ell$, $\Pi_\nu \ni P_{\varepsilon_s}(s_\ell)$ とすると、意味解釈オペレータ S_p は、T_ℓ から Π_ν への作用素として定義される。また、$\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_\ell\}$ は、特徴付ベクトルであり、データ行列 A の特徴と同一の特徴を用いている。

オペレータ S_p は次のように定義される。

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

   $\mathbf{u}_i$ と $\mathbf{q}_j$ の内積を u_{ij} とする。

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

   ベクトル $\hat{\mathbf{u}}_i \in \mathcal{I}$ を次のように定める。

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

   これは、単語 $\mathbf{u}_i$ をイメージ空間 $\mathcal{I}$ に写像したものである。

2. 文脈語群 s_ℓ の意味重心 $\mathbf{G}^+(s_\ell)$ を求める。

$$\mathbf{G}^+(s_\ell) := \frac{\left(\sum_{i=1}^{\ell} u_{i1}, \sum_{i=1}^{\ell} u_{i2}, \cdots, \sum_{i=1}^{\ell} u_{i\nu}\right)}{\| \left(\sum_{i=1}^{\ell} u_{i1}, \sum_{i=1}^{\ell} u_{i2}, \cdots, \sum_{i=1}^{\ell} u_{i\nu}\right) \|_\infty}$$

この $\|\cdot\|_\infty$ は、無限大ノルムを示す。

3. 意味射影 $P_{\varepsilon_s}(s_\ell)$ の決定

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

但し $\Lambda_{\varepsilon_s} := \{\, i \,|\, (\mathbf{G}^+(s_\ell))_i > \varepsilon_s \}$ とする。

2.2.4 意味空間における距離計算

単語 $\mathbf{x}$ と単語 $\mathbf{y}$ 間の距離 $\rho(\mathbf{x}, \mathbf{y}; s_\ell), \mathbf{x}, \mathbf{y} \in \mathcal{I}$ を次のように定める。

$$\rho(\mathbf{x}, \mathbf{y}; s_\ell) = \sqrt{\sum_{j \in \Lambda_{\varepsilon_s}} \{c_j(s_\ell)(x_j - y_j)\}^2},$$

ここで、$c_j(s_\ell)$ は、文脈語群 s_ℓ に依存して決まる重みであり、次のように定義する。

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

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

3 基本モデルの拡張

3.1 静的意味識別オペレータ

イメージ空間を構成する意味素を作成するとき、単語の分布に偏りのある意味素 (以下、“主軸” と呼ぶ) ができるため、適切な単語間の関係の解釈が行われないことがある。主軸ができる原因は、固有値分解を行うとき、主成分分析と等価な方法により意味素を求めているためと考えられる。この方法は、単語の分散が高い順に意味素を決定する。そのため、単語の分布に偏りのある主軸ができることがある。主軸は、どのような文脈語群においても、意味重心との相関が高くなりやすく、意味射影の対象の空間に含まれる可能性がある。そのため、主軸上における単語間の関係が解釈に影響し、適切な解釈が行われないことがある。このような場合、文脈に応じて意味解釈オペレータによって構成された意味射影から、主軸への射影を排除する必要がある。そこで、意味射影から主軸への射影を排除した射影 (以下、“意味識別射影” と呼ぶ)$D_{\varepsilon_{ds}}$ を次のように求める。

まず、データ行列に登録されている全ての単語 m 個に対する特徴付ベクトルをイメージ空間 $\mathcal{I}$ へ写像したベクトル

$$\hat{\mathbf{u}}_i \ , \ \hat{\mathbf{u}}_i \in \mathcal{I} \ , \ i = 1, 2, \cdots, m$$

と、正数 $\varepsilon_{ds} (\varepsilon_{ds} > 0)$ が与えられたとする。

また、$\hat{\mathbf{u}}_i$ の和のベクトルを

$$\begin{aligned} \mathbf{t} &:= \left(\sum_{i=1}^{m} u_{i1}, \sum_{i=1}^{m} u_{i2}, \cdots, \sum_{i=1}^{m} u_{i\nu}\right) \\ &:= (t_1, t_2, \cdots, t_\nu) \end{aligned}$$

とする。そして、ベクトル $\mathbf{t}$ の要素において、x 番目に絶対値の大きな要素の添字を求める関数を $A\_MAX(\mathbf{t}, x)$ とする。

このとき、添字集合 $\Lambda_{\varepsilon_{ds}}$ を次のように求める。

1. 添字集合 $\Lambda_{\varepsilon_{ds}}$ を空集合に初期化する。
2. ループ変数 i を 1 から $\nu - 1$ まで、次の 3 から 6 を繰り返す。
3. 添字変数 j の値を $A\_MAX(\mathbf{t}, i)$ とし、添字変数 k の値を $A\_MAX(\mathbf{t}, i+1)$ とする。
4. $\log_e \frac{|t_j|}{|t_k|} < \varepsilon_{ds}$ ならば、ループを抜けて、終了する。
5. 添字集合 $\Lambda_{\varepsilon_{ds}}$ に添字変数 j の値を加える。
6. ループ変数 i に 1 を加算し、3 へ行く。

そして、意味識別射影 $D_{\varepsilon_{ds}}$ を次のように定義する。

$$D_{\varepsilon_{ds}} := \sum_{i \in \Lambda_{\varepsilon_s} \setminus \Lambda_{\varepsilon_{ds}}} P_{\lambda_i} \in \Pi_\nu$$

また、静的意味識別オペレータを考慮した、文脈語群 s_ℓ における単語 $\mathbf{x}$ と単語 $\mathbf{y}$ 間の距離 $\rho(\mathbf{x}, \mathbf{y}; s_\ell), \mathbf{x}, \mathbf{y} \in \mathcal{I}$ を次のように定める。

$$\rho(\mathbf{x}, \mathbf{y}; s_\ell) = \sqrt{\sum_{j \in \Lambda_{\varepsilon_s} \setminus \Lambda_{\varepsilon_{ds}}} \{c_j(s_\ell)(x_j - y_j)\}^2}$$

3.2 イメージ空間へのキーワードの写像の方式

意味重心の符号を考慮せずに解釈の対象となるキーワード (以下、“検索キーワード” と呼ぶ) をイメージ空間へ写像した場合、文脈に応じた単語間の関係の解釈が正しく行われないことがある。その例を図 2に示す。まず、データ行列として、図 2(a) が与えられたとする。そして、データ行列の空間とイメージ空間が、図 2(b) の位置関係にあるとする。このとき、検索キーワードとして “computer”、距離計算の対象となるキーワード (以下、“比較対象語” と呼ぶ) として “software” と “hardware”、文脈語として “software” が与えられた場合について考える。文脈語が “software” のため、その文脈が示す意味空間は、意味素 $\mathbf{q}_1, \mathbf{q}_2$ によって張られる空間になる。その意味空間において、検索キーワードと各比較対象語の距離計算をすると、文脈語が “software” にも関わらず、図 2(c) に示すように、“computer” から等距離上に “software” と “hardware” がある。これは、文脈語が “software” にも関わらず、解釈ができていないことを意味する。

この原因は、文脈が示している意味に対し、反対の意味が写像後の検索キーワードに含まれているためと考えられる。文脈に反対の意味を含めて写像した場合、本来、意味を識別するために重要な意味素において、特徴が相殺されてしまう。そのため、解釈が正しく行われないことがある。

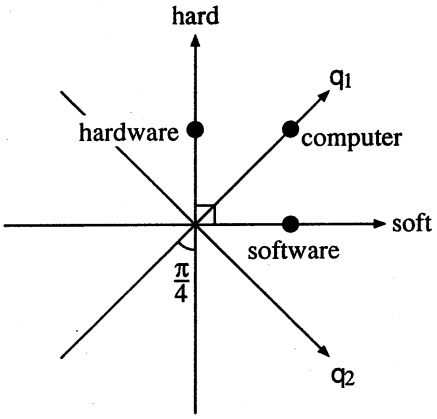
そこで、イメージ空間への検索キーワードの写像では、検索キーワードから文脈と関係のない要素を取り除くことが必要であると考えられる。その方法として、以下のように写像を行う。

1. まず、検索キーワードの特徴付ベクトル

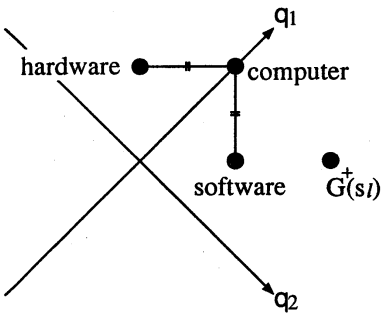
$$\mathbf{u} := (u_1, u_2, \cdots, u_n)$$

単語 \ 特徴	hard	soft
computer	1	1
hardware	1	0
software	0	1

（a）データ行列

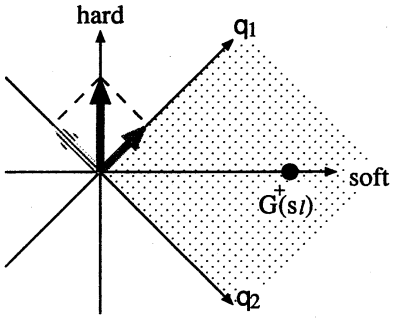


（b）データ行列の空間と
イメージ空間

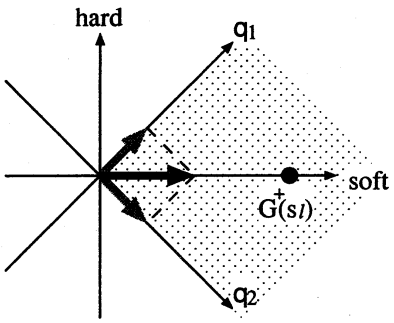


（c）距離計算

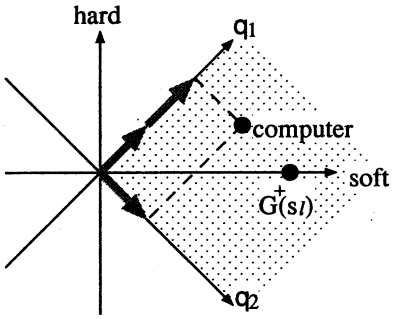
図 2: 意味重心の符号を考慮しないキーワードの写像と
その場合の距離計算



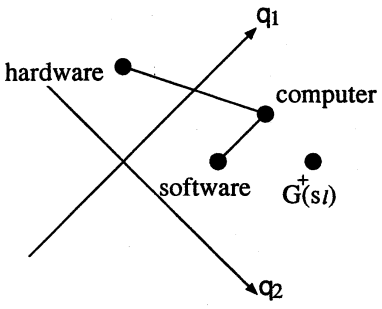
（a）特徴ごとの写像（h a r d）



（b）特徴ごとの写像（s o f t）



（c）特徴ごとに文脈を考慮して
写像した結果



（d）距離計算

図 3: 意味重心の符号を考慮したキーワードの写像と
その場合の距離計算

を、次のように各特徴ごとに分解する。

$$\mathbf{u}_1^{'} := (u_1, 0, \cdots, 0)$$

$$\mathbf{u}_2^{'} := (0, u_2, \cdots, 0)$$

$$\vdots$$

$$\mathbf{u}_n^{'} := (0, 0, \cdots, u_n)$$

2. 分解したベクトル $\mathbf{u}_i^{'}$ を写像する。

   分解したベクトル $\mathbf{u}_i^{'}$ と意味素 $\mathbf{q}_j$ の内積を $u_{ij}^{'}$ とする。但し、内積の値 $u_{ij}^{'}$ が意味重心の要素 g_j と異符号のときは、内積の値 $u_{ij}^{'}$ を "0" とする。

$$u_{ij}^{'} := \begin{cases} (\mathbf{u}_i^{'}, \mathbf{q}_j), & for\ (\mathbf{u}_i^{'}, \mathbf{q}_j) \cdot g_j >= 0 \\ 0, & for\ (\mathbf{u}_i^{'}, \mathbf{q}_j) \cdot g_j < 0 \end{cases}$$

$$G^{+}(s_l) := (g_1, g_2, \cdots, g_\nu)\ ,\ j = 1, 2, \cdots, \nu$$

$$,\ i = 1, 2, \cdots, n$$

3. イメージ空間における検索キーワードのベクトル $\hat{\mathbf{u}} \in \mathcal{I}$ を次のように定める。

$$\hat{\mathbf{u}} := \left(\sum_{i=1}^{n} u_{i1}^{'}, \sum_{i=1}^{n} u_{i2}^{'}, \cdots, \sum_{i=1}^{n} u_{i\nu}^{'} \right)$$

この操作を行った例を、図3に示す。この例では、前例と同じキーワードと空間を使用している。まず、検索キーワードの特徴付ベクトルを各特徴ごとに分解し、写像した例を図3(a)(b)に示す。図3(a)は、特徴 "hard"、また、図3(b)は、特徴 "soft" について、写像している。このとき、図3(a)において、意味素 $\mathbf{q}_2$ との内積の値は、意味重心の符号と異なる。そのため、その内積の値を、写像後の検索キーワードのベクトルから取り除く。その結果、検索キーワードのベクトルは、イメージ空間において図3(c)になる。また、各比較対象語との距離は図3(d)になり、"software" の方が "hardware" より近くなる。これは、意味重心の符号を考慮した検索キーワードの写像を行うことにより、文脈に応じた単語間の関係の解釈が正しく行われたことを意味する。

4 実現

4.1 意味空間におけるキーワード間の距離計算の方法

ある文脈語群が与えられ、検索キーワードに最も意味の近い比較対象語を求めるとき、全ての比較対象語との距離を計算していたのでは、利用者に対し素早い対応ができない。そこで、次の順序に従って距離計算を行うことにより、全ての比較対象語との距離計算をしなくても、検索キーワードに最も意味の近い比較対象語を探すことができる。

1. 検索キーワードのベクトルを

$$\mathbf{x} := (x_1, x_2, \cdots, x_\nu) \ , \ \mathbf{x} \in \mathcal{I}$$

とし、比較対象語群のベクトルを

$$\mathbf{y}_i := (y_{i1}, y_{i2}, \cdots, y_{i\nu}) \ , \ \mathbf{y}_i \in \mathcal{I}$$
$$, i = 1, 2, \cdots, m$$

とする。

2. 意味重心 $\mathbf{G}^+(s_l)$ と最も相関がある意味素を $\mathbf{q}_j$ とする。

3. 範囲変数 Δ_h を ∞ に初期化する。また、検索キーワードとの距離が最も近い比較対象語の候補 $\mathbf{z}$ を $NULL$ に初期化する。

4. 検索キーワードとの距離を計算していない比較対象語のベクトルの集合 Y を求める。このとき、集合 Y が空集合ならば、比較対象語 $\mathbf{z}$ を選択して終了する。

5. 集合 Y の要素から、意味素 $\mathbf{q}_j$ 上において、検索キーワードのベクトルの要素 x_j に最も近い要素 y_{kj} $(1 \leq k \leq m)$ を持つ比較対象語 $\mathbf{y}_k \in Y$ を探す。

6. 意味素 $\mathbf{q}_j$ 上において $x_j \pm \Delta_h$ の範囲内に y_{kj} が含まれないなら、比較対象語 $\mathbf{z}$ を選択して終了する。

7. 検索キーワードのベクトル $\mathbf{x}$ と比較対象語のベクトル $\mathbf{y}_k$ との距離 $\rho(\mathbf{x}, \mathbf{y}_k)$ を求める。

8. もし、範囲変数 Δ_h より距離 $\rho(\mathbf{x}, \mathbf{y}_k)$ の方が小さいならば、範囲変数 Δ_h を $\rho(\mathbf{x}, \mathbf{y}_k)$ とし、検索キーワードとの距離が最も近い比較対象語の候補 $\mathbf{z}$ をベクトル $\mathbf{y}_k$ の比較対象語とする。

9. 4 へ行く。

以上の処理を実現するために、あらかじめ、各単語をイメージ空間へ写像した行列を用意しておき、さらに、各意味素ごとに、その意味素における各単語の値をソートした行列を用意する。これらの行列と前に述べた処理により、意味空間において検索キーワードに最も意味の近い比較対象語を求めるとき、距離計算の回数を減らすことができる。

4.2 データ行列の自動生成の方法

本モデルの実現において、データ行列を自動生成するために英々辞典をもちいた。英々辞典には、限られた基本語のみを使用して、説明文や例文を書いているものがある [16][17][18]。その基本語をデータ行列の特徴と一致させ、見出し語をデータ行列の各単語に対応させることにより、自動生成が可能となる。この自動生成において、データ行列の各要素の値は、見出し語の説明文中に基本語が肯定の意味にもちいられていた場合 “1”、否定の場合 “-1”、使用されていない場合、 “0” とした。そして、イメージ空間の作成用のデータ行列とその空間へ写像する単語群のデータ行列を、それぞれに適していると考えられる方法により正規化した。この正規化の有効性については、次章において述べる。

また、英々辞典からデータ行列を自動生成するためのフィルタ群を作成した。各フィルタの機能を次に示し、そのフィルタ群をもちいてデータ行列を作成していく過程の例を図 4に示す。

辞書テキスト:

ditch, 1. n. Narrow waterway for draining fields, roads. 2. v.i.&.t. Go,...

street, n. Road lined with buildings. *Not in the same s. as,...*

filter-1

ditch, 1. n. narrow waterway for draining fields, roads.

street, n. road lined with buildings.

filter-2

ditch narrow waterway for draining fields roads street road lined with buildings

filter-3

ditch narrow water way for draining fields roads street road lined with buildings

filter-4

ditch narrow water way draining fields roads street road lined buildings

filter-5

ditch narrow water way drain field road street road line building

filter-6

作成されたデータ行列

単語＼特徴	···field···line···road···way···
⋮	··· ··· ··· ··· ··· ··· ··· ···
ditch	··· 0 ··· 0 ··· 1 ··· 1 ···
⋮	··· ··· ··· ··· ··· ··· ··· ···
street	··· 1 ··· 1 ··· 1 ··· 0 ···
⋮	··· ··· ··· ··· ··· ··· ··· ···

図 4: データ行列の作成過程

Filter-1 辞典から、必要な見出し語とその説明文を切り出す。その際、大文字を小文字へ変換する。

Filter-2 特別な記号 (セミコロン、コンマ等) と品詞を削除する。

Filter-3 合成語を複数の基本語に分解する。

Filter-4 意味の識別に必要のない単語 (冠詞、 be 動詞、代名詞、間投詞、接続詞、前置詞、助動詞) を除去する。

Filter-5 語尾変化している単語を基本語へ変換する。

Filter-6 以上のフィルタによって得られた単語群をもとに、データ行列へ変換する。

5 結論

本稿では、意味の数学モデルの実現方式について述べ、さらに、その方式によって実現した実験システムを用いて実験を行い、その結果から、単語間の意味的な同一性、相異性の判定に、本モデルが有効であることを明らかにした。今後は、実在するデータベース・システムの意味的な同一性、相異性を扱う場合の問題において、本モデルの有効性を示す必要があると考えている。

参考文献

[1] Batini, C.,Lenzelini, M. and Nabathe, S.B., *"A comparative analysis of methodologies for database schema integration,"* ACM Comp. Surveys, **18**, pp.323-364, 1986.

[2] Fang, D., Hammer, J., Mcleod, D., *"The identification and resolution of semantic heterogeneity in multidatabase systems,"* Proc. 1st Int. Workshop on Interoperability in Multidatabase Systems, pp. 136-143, Apr. 1991.

[3] Gallant, S.I., *"A practical approach for presenting context and for performing word sense disambiguation using neural networks,"* Neural Computation, **3**, pp.293-309, 1991.

[4] Pu, C., *"Semantic based integration library: A proposal for cooperative research for semanitic interoperability,"* Proc. Workshop on Multidatabases and Semantic Interoperability, pp6-9, Nov., 1990.

[5] T. Kitagawa and Y. Kiyoki, "A mathematical model of meaning and its application to multidatabase systems," Proceedings of 3rd IEEE International Workshop on Research Issues on Data Engineering: Interoperability in Multidatabase Systems, pp.130-135, April 1993.

[6] Y. Kiyoki and T. Kitagawa, "A metadatabase system for supporting semantic interoperability in multidatabases," Information Modelling and Knowledge Bases (IOS Press), Vol. V, pp.287-298, 1993.

[7] Y. Kiyoki and T. Kitagawa, "A semantic associative search method for knowledge acquisition," Information Modelling and Knowledge Bases (IOS Press) (to be published), Vol. VI, 1995.

[8] T. Kitagawa and Y. Kiyoki, "A new information retrieval method with a dynamic context recognition mechanism," Proceedings of 47th Conference of International Federation for Information and Documentation, pp.210-215, Oct. 1994.

[9] Y. Kiyoki, T. Kitagawa and Y. Hitomi, "A fundamental framework for realizing semantic interoperability in a multidatabase environment," International Journal of Integrated Computer-Aided Engineering (John Wiley & Sons), **2**, pp.3-20, 1995.

[10] Y. Kiyoki and T. Hayama, "The design and implementation of a distributed system architecture for multimedia databases," Proceedings of 47th Conference of International Federation for Information and Documentation, pp. 374-379, Oct. 1994.

[11] Y. Kiyoki, T. Kitagawa and T. Hayama, "A Metadatabase System for Semantic Image Search by a Mathematical Model of Meaning, ACM SIGMOD Record,ACM SIGMOD Record, Vol.23, No. 4, pp.34-41, Dec. 1994.

[12] Sheth, A. and Larson, J.A., *"Federated database systems for managing distributed, heterogeneous, and autonomous databases,"* ACM Comp. Surveys, **22**, pp183-236, 1990.

[13] Sheth, A. and Kashyap, V., *"So far (schematically) yet so near (semantically),"* Proc. IFIP TC2/WG2.6 Conf. on Semantics of Interoperable Database Systems, Nov. 1992.

[14] Shimizu, H., Kiyoki, Y., Sekijima, A. and Kamibayashi, N., *"A Decision Making Support System for Selecting Appropriate Online Databases,"*, Proc. 1st Int. Workshop on Interoperability in Multidatabase Systems, pp.322-329, Apr. 1991.

[15] Yu, C., Sun, W., Dao, S., Keirsey, D.,
"Determining relationships among attributes for interoperability of multi-database systems," Proc. Workshop on Multidatabases and Semantic Interoperability, pp10-15, Nov. 1990.

[16] Ogden, C.K., *"The General Basic English Dictionary,"* Evans Brothers Limited, 1940.

[17] *"Longman Dictionary of Contemporary English,"* Longman, 1987.

[18] Hill, L.A., *"Penguin English Student's Dictionary,"* Penguin Books Limited, 1991.