意味の数学モデルによる高速な意味的連想処理方式 *

3G－4

宮原 隆行　清木 康 †
筑波大学 電子・情報工学系 ‡

1 序論

近年、広域ネットワークを介して、多種多様なデータベースの利用が可能になっている。マルチデータベース・システムにおける重要な課題の一つは、異なるデータベース間の単語間の意味の同一性、相異性の扱いである。多種多様なデータベース群を統一的に扱うマルチデータベース・システムを構築するためには、データの表す意味を扱うことが重要になる。このような概念はセマンティック・インタオペラビリティと呼ばれる。

現行のデータベース・システムにおいては、単語間の意味的な同一性、相異性に関する関係を静的かつ明示的に記述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。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 [1, 2, 3]。

意味の数学モデルは、ある単語に意味的に近い単語を大量のデータの中から、高速に抽出する能力を潜在的に備えている。本稿では、意味の数学モデルにおける高速な情報獲得 (高速意味的連想処理) の実現方法を示す。

2 意味の数学モデル

1. **前提**：まず、単語の意味を決定する語として**特徴**を設定する。単語の意味をいくつかの特徴で表すことを**特徴づけ**という。m 個の単語があり、各単語は n 個の特徴で特徴づけされる。これをデータ**行列**と呼び、本モデルに対してこのデータ行列が与えられているものとする (図 1)。

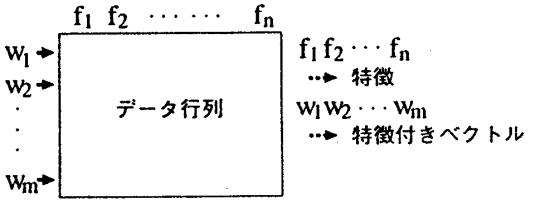

図 1: データ行列の構成

2. **イメージ空間 I の設定**：データ行列から、特徴づけに関する相関行列を作る。相関行列を固有値分解し、固有ベクトルを正規直交化する。非ゼロ固有値に対応する固有ベクトル (以下、**軸**と呼ぶ) の張る空間を**イメージ空間**と定義する (図 2の左図)。このとき任意の言葉は、イメージ空間上へ写像でき (言葉のフーリエ変換)、その空間内の一点となる (図 2の右図)。この空間の次元 ν は、データ行列のランクに一致する。このような空間設定は、相関行列の対称性からいつも可能である。またこの空間は、ν 次元ユークリッド空間となり、様々なノルムが入る。

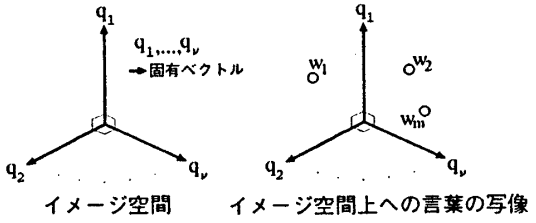

図 2: イメージ空間の設定

3. **意味射影の集合 Π の設定**：イメージ空間の固有部分空間は、いくつかの軸の組み合わせである。これを**部分空間**と呼ぶ。**意味射影**とは、言葉を空間へ射影することをいう。ここでは、イメージ空間上の 1 点 (言葉) に対する部分空間への全ての意味射影の集合を考える。i 次元の部分空間は、$(\nu(\nu-1)\cdots(\nu-i+1))/i!$ 個存在するので、意味射影の総数は 2^ν 個となる。すなわち本モデルは、2 の ν 乗通りの意味の様相の表現能力を持つ。

4. **意味解釈オペレータ S_p の構成**：文脈を決定する ℓ 個の言葉の列 s_ℓ が与えられたとき、その文脈に応じた意味射影を以下のように決定する。

   a) 1 つの言葉をイメージ空間内でフーリエ展開し、フーリエ係数を求める。これは、その言葉と各軸との相関を求めることに相当する。この操作を ℓ 個の単語について行う。

   b) 各軸毎に、a) で求めたフーリエ係数の絶対値の総和を求める。これは、言葉の列と各軸との相関を求めることに相当する。

   c) 与えられたしきい値 ε に対して、b) で求めた総和が大きくなる軸がある場合、その軸に対応する部分空間への射影を**意味解釈オペレータ S_p** の値とする。

---

*Fast Semantic Associative Search by the Mathematical Model of Meaning

†Takayuki Miyahara and Yasushi Kiyoki
‡Institute of Information Sciences and Electronics,University of Tsukuba

5. **部分空間における距離計算**：言葉は、ν 次元ユークリッド空間上における一点で表される。したがって言葉と言葉の意味的関係は、ユークリッド距離として計算できる。この距離が短い場合は「意味が近い」、長い場合は「意味が遠い」と判断する。ただし、部分空間を構成する各軸は、その文脈における重みを持ち、その重みを各軸上の言葉のフーリエ係数に乗算している。これは、文脈の持つ意味 (ここでは、各軸の重み) を距離計算に反映させるためである。

3 意味的連想処理

意味の数学モデルにおける意味的連想処理とは、文脈を表す単語列 (文脈語群と呼ぶ) によって選ばれた意味空間の中で、ある単語 (キーワードと呼ぶ) に最も近い意味の単語を、単語群 (比較対象語群と呼ぶ) の中から選ぶことである。

3.1 高速意味的連想処理方式

意味の数学モデルを使用した高速な意味的連想処理を実現するアルゴリズムを提案する。提案するアルゴリズムでは、意味的連想処理の計算のオーダは、$O(n)$ から $O(1)$ へ向かって減少する。

3.2 アルゴリズム

意味の数学モデルでは、重みの付いた高次元のユークリッド空間において、キーワードと比較対象語との距離計算を行い、キーワードとの距離が最も短い比較対象語を求めることによって、意味的連想処理を行う。この時の処理において、キーワードとの距離の長い比較対象語を随時処理範囲から除くことによって、意味的連想処理の高速化を行う。前提として、あらかじめ、全比較対象語群を、各軸ごとに座標点順にソートされているものとする。以下に、具体的なアルゴリズムを示す。

(1) 文脈にしたがって選択された部分空間内において、重みの最も重い軸上で、キーワードに最も近い比較対象語を選び出す。(図 3 内の 1)

(2) キーワードと選び出した比較対象語との間で、部分空間内における距離を求める。(図 3 内の 2)

(3) 最も重みのある軸上で、(2) において求めた距離と等しい長さに対応する点を求め、マークする (図 3 内の 3)。そのマークよりも遠い位置にある比較対象語群を処理対象から排除する。

(4) 最も重みのある軸において、キーワードとマークの間に存在している比較対象語群だけを処理対象とし、それぞれについて、部分空間におけるキーワードとの距離を計算し、キーワードに最も近い距離をもつ比較対象語を解とする。

((4) において、キーワードとマークの間に存在している比較対象語群の中で、最も重みのある軸において、(1) において選んだ比較対象語を除いて、キーワードに近い順に比較対象語選ぶ (図 3 内の 4)。そして、(2) を行った後、その距離が現在のマークとキーワード間の距離よりも短い場合には、(3) の操作により、その点を新たなマークとして設定し、そのマークよりも遠い位置にある比較対象語をさらに排除する。これにより、比較対象語の数をさらに減らしていくことができる。)

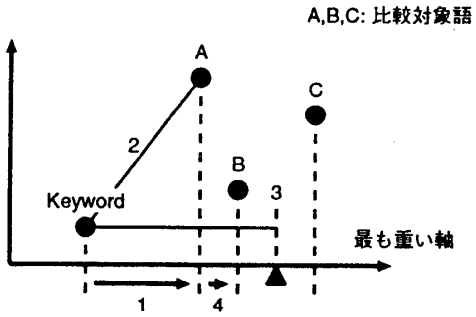

図 3: 高速化のアルゴリズム

このように、処理範囲を狭めることにより、意味的連想処理において、距離計算を行なわねばならない比較対象語の数を減らすことが可能となり、意味的連想処理の高速化を実現できる。また、本アルゴリズムを適用した場合、意味的連想処理は、処理範囲を狭めることが一度もできない場合に、距離計算回数は $O(n)$ になり、隣接する一単語の距離を調べただけで、他の単語の距離を計算する必要がなくなる場合に、距離計算回数は $O(1)$ になる。

なお、特定の軸上で、キーワードに近い単語を順次求める処理については、あらかじめ各軸ごとに、すべての単語を座標順にソートしておくという処理を行なっているので、高速に実現できる。

3.3 まとめ

意味の数学モデルにおける、高速意味的連想処理方式を提案した。今後は、学習機構との組合せを行ない、より高速な意味的連想処理を実現していく予定である。

参考文献

[1] 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.

[2] 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) (to be published), 1995.

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