第 2 章

意味の数学モデル

2.1 概要

ここでは，意味の数学モデルの概要を示す．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．

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 を超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 定式化

本節では，意味の数学モデル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 $\mathrm{w}_i (i = 1, \cdots, m)$ 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．

1. データ行列 A の相関行列 A^TA を作る．

$$
\begin{array}{c|ccc|}
 & \mathbf{f}_1 & \mathbf{f}_2 \quad \cdots & \mathbf{f}_n \\
\hline
\mathbf{w}_1 \rightarrow & & & \\
\mathbf{w}_2 \rightarrow & & & \\
 & & A & \\
\vdots & & & \\
\mathbf{w}_m \rightarrow & & & \\
\hline
\end{array}
$$

2. $A^T A$ を固有値分解する．

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

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

ここで行列 Q は，

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

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．

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

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

意味射影の集合 Π_ν を次のように定義する．

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

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 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 $\widehat{\mathbf{u}}_i \in \mathcal{I}$ 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 $\Lambda_{\varepsilon_s} := \{\, i \mid (\mathbf{G}^+(s_\ell))_i > \varepsilon_s \}$ とする．

2.2.4 意味空間における距離計算

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

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

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.

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

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

2.3 意味の数学モデルによる連想検索の実現

意味空間における距離が定まると, 文脈語を与え, 検索キーワードを与えることで, 検索対象単語群の中から, 検索キーワードに最も近い単語を選び出すことが可能となる. この動作が, 意味の数学モデルによる連想検索である.

意味の数学モデルによる連想検索システムの実現にあたり, “The General Basic English Dictionary”[24] と “Longman Dictionary of Contemporary English”[22] の 2 つの英英辞典を用いた. これらの辞書は基本単語が決められており, 基本単語のみで全ての単語を説明している.

まず, 最初に “Longman Dictionary of Contemporary English” における約 2,000 の基本英単語について, “The General Basic English Dictionary” の説明語を参照した. ただし, 基本語の活用形なども説明語に使用されているので, それらの単語を行列作成にする場合において, 次のフィルター群を用いて, 全ての単語の原型を使用した.

1. 熟語や慣用表現を同義語 1 単語に変換
2. 辞書中で「基本語以外の単語」とされている語を削除
3. 基本語に接頭語, 接尾語, 他の基本語が接続している語を基本語 (の列) に変換
4. 基本語の活用形や複数形を原型や単数形に変換
5. 未定義の基本語以外の単語 (固有名詞など) を削除
6. 句読記号, 文法記号など, 不要な記号を削除
7. 特徴語とするのに不適当な基本語を削除

8. 合成する必要のある基本語を合成した形に変換

上記の様に単語の原型のみの状態にした後，説明語の中に基本語が出ていれば 1，出ていなければ 1，反対の意味に用いられていれば -1 を与え，2,000 × 800 の行列を作成した．次に行列のトランスポーズと行列そのものを掛け，固有値分解を行い，得られた固有ベクトルのみ構成される意味空間を作成した．検索で用いられる単語群も，フィルター群を用いて 0, 1, -1 の状態にし，ベクトル群を構成した．

その後，文脈や状況を与えるベクトルによって部分空間を切り出し (閾値 εs を超えた軸のみで部分空間を構成する)，検索対象語や，キーワードとなる点を，その空間に乗せ，キーワードに最も近い検索対象語を，文脈や状況に応じた検索対象語として取り出す．この一連の操作が意味の数学モデルによる連想検索である [32]．

本モデルを利用すると，記号ではなく，意味による情報伝達が可能になる．パターンマッチでは不可能なことが可能になるのである．実際，本システムを利用して，マルチデータベースシステムへの応用研究 [12, 14, 16, 29, 34]，基本システムの高速化研究 [33, 48]，画像データベースへの適用研究 [49, 30, 44]，学習システムの研究 [34]，メディア情報検索システムの実現 [50] の研究がなされている．特に同一メディア間ではなく，異種メディアによる検索が可能であることを強調したい．画像データベースシステムへの適用研究では，画像を言葉で検索している．言葉で表せるものであれば，全て本モデルで扱うことが可能なのである．