「言語基底の考察」

□「言語空間」を位相空間とみなし、言語の定義を位相上で解説するには数学的な考え方から順を追ってしなくてはならず、たいへん解り難いので、まず解り易い距離空間への定義を説明してからにする。言語空間を距離空間へ線形性とノルム、そして完備性と内積を導入した「ヒルベルト空間」と定義し、単語$\{x_i\}$をヒルベルト空間Hにおける一次独立なベクトルの有限列か可算列とみなせば、正規直交系$\{e_i\}$が存在して$[\{x_i\}]=[\{e_i\}]$が「グラム‐シュミットの直交化定理」により成り立つ。

簡単に証明を書けば、

■■　$e_i=\dfrac{x_i}{\|x_i\|}$と定義し、$e_n(n\geq 2)$を次の漸化式によって定義する。

$$z_n=x_n-\sum_{i=1}^{n-1}(x_n\,|\,e_n)e_n,\quad e_n=\frac{z_n}{\|z_n\|}$$

このとき、$\|e_n\|=1\ (\forall n)$は明らかである。$\{e_i\}$の直交性について、

$(z_2\,|\,e_1)=(x_2-(x_2\,|\,e_1)e_1\,|\,e_1)=(x_2\,|\,e_1)-(x_2\,|\,e_1)=0$　より、$(e_2\,|\,e_1)=0$。
いま、$(e_n\,|\,e_k)=0\quad(k=1,\cdots,n-1)$と仮定すると、

$$(z_{n+1}\,|\,e_k)=(x_{n+1}\,|\,e_k)-\sum_{i=1}^{n}(x_{n+1}\,|\,e_i)(e_i\,|\,e_k)=(x_{n+1}\,|\,e_k)-(x_{n+1}\,|\,e_k)=0$$

また、$(z_{n+1}\,|\,e_n)=0$
帰納法によって、
$\forall n;(e_n\,|\,e_k)=0\quad(k-1,\cdots,n-1)$が成り立つから、$\{e_i\}$は正規直交系である。

また、$[e_1]=[x_1]$は明らかであり、$[e_1,\cdots,e_n]=[x_1,\cdots,x_n]$とすると、

$e_{n+1}\in[e_1,\cdots,e_n,x_{n+1}]$より、
$$[e_1,\cdots,e_n,e_{n+1}]=[e_1,\cdots,e_n,x_{n+1}]=[x_1,\cdots,x_n,x_{n+1}]$$
ゆえに、$\forall n;;[e_1,\cdots,e_n]=[x_1,\cdots,x_n]$だから$[\{x_i\}]=[\{e_i\}]$。　■■

すなわち、言語空間上に「単位ベクトル」と「軸の直交性」が証明されたわけです。言い換えれば、言語空間を「ヒルベルト空間」と定義すれば、「基底」と「直交性」が保証されるということで、「内積」やその他の便利な手法が使えることになる。

そこで、
上記言語空間であるヒルベルト空間の「近似部分空間」が有限個の基底$y_1,\cdots,y_n$で生成されていて、おのおのが互いに直交し大きさが１であれば、正規方程式の左辺の行列が単位

行列となり、任意のベクトルxの近似は、$\sum_{i=1}^{n}(x \mid y_i)y_i$と簡単に計算できる。従って、近似部分空間をこのように正規化された基底で生成できれば、近似問題は解け易くなる。

本来は、このような証明をしてから、単語の類似関係に余弦定理や内積、共分散分析、Nグラムや隠れマルコフ理論などを使わねばならない。これらを定義せずに安易に数理的手法を使うのは、利用範囲と精度の限界とで、とんでもない結果を招いたり、精度を収束させるときに新しい改善法を発見できないことになる。次はTopologyの中での「基底」について考察する。(第1版)

⇒ *cTag*＞意味位相空間ページへ