日本知能情報ファジィ学会誌「用語解説」(2005 年 2 月号掲載予定)

潜在意味解析(Latent Semantic Analysis;LSA)

潜在意味解析とは,さまざまな文脈において語の意味がどのように使用されているかを大規模なテキストコーパスに現れるすべての語や語の集合(文,小論など)に対し統計的な計算(特異値分解)と次元縮約を施すことによって導出し意味空間として表現する理論である.この理論の根幹をなす考え方は,文脈によって特定の語が現れたり現れないということが,語と語,あるいは語の組み合わせの間の意味的な類似性を決定するというものである.意味空間は以下の手続きで構成される.まず,テキストコーパスが与えられたとき,ある語がある文脈で現れる頻度を適当に変換して得られる行列 A(n 語 $\times m$ 文脈,階数 r)を作成し,$\underset{n\times m}{A} = \underset{n\times r}{U}\,\underset{r\times r}{D_\lambda}\,\underset{r\times m}{V^T}$ の形に分解する(特異値分解).D_λ は $\lambda_1 \geq \ldots \geq \lambda_r > 0$(特異値)を対角要素にもつ対角行列,$U$,$V$ は正規直交ベクトルを列ベクトルにもつ行列 $(U^TU = V^TV = I)$ である.$U = (u_1, \ldots, u_r)$,$V = (v_1, \ldots, v_r)$ とおけば,$A = \sum_{i=1}^{r} u_i\lambda_i v_i^T$ と表せる.u_i,v_i は左あるいは右特異ベクトルと呼ばれる.次に,行列 A の階数を r から k に減じることにより $A_k = \sum_{i=1}^{k} u_i\lambda_i v_i^T$ を作成する.この行列は,階数を k とした場合の行列 A の最適近似となることが知られている.それぞれの語や文脈は,左,右特異ベクトルの k 個の要素を用いて k 次元空間のベクトルとして表現される.次元縮約後の行列 A_k においては,もとの行列 A に含まれている語の使用法が多様であることに起因するノイズが除かれ,意味的に近い語が近い場所に配置させられている.米国大学生の言語的知識に対応する意味空間が作成されているが,その場合,$n = 92,409$,$m = 37,651$,$k = 419$ である.このようにして作成される意味空間が人間の知識を反映していることは,さまざまな方面で検証されている.たとえば,潜在意味解析によって作成された意味空間を用いて語彙テストや読解テストを行ったとき,人間の成績と似たパタンを示すことが報告されている.

(産業技術総合研究所　人間福祉医工学研究部門　北島　宗雄)