3L-5

情報検索システムのための，単語の意味の空間的表現と学習

堀　浩一　　戸田誠之助　　安永尚志
(国文学研究資料館　情報処理室)

1．まえがき

従来，論文検索システムをはじめとする情報検索システムのために，所望の情報にたどりつくためのキーワードの抽出法と使用法について，さまざまな研究がなされてきた．キーワードの構造化，キーワードの種類のコントロール，シソーラスの作成法と利用法などが考えられてきたが，結局のところ，大規模なシステムにおいては，単純な，構造のない，また，コントロールされないフリーキーワードが使われることが多かった．その理由は，システム製作者がどんなにがんばっても，構造や，キーワードの種類を把握しきれないためであると考えられる．これは，自然言語の意味処理全般に通じる問題であって，どんなに工夫しても，意味のあらゆる側面をカバーする表現体系をあらかじめ与えることは不可能である．

そこで，筆者らは，ユーザ1人1人に合わせて，単語の意味を学習するシステムを考えた．単語の意味は，単語と概念からなる空間の構造として表現される．意味を空間的に表現しようというアイディアそのものは新しいものではないが，ユーザに合わせて空間を変形していくというメカニズムを与えることにより，あらかじめ把握できない意味も学習されるようになり，単純なキーワードでは検索不可能だった情報を，意味の空間的表現を介して，見付けることが論文検索システムにおいて可能になった．文学の論文検索システムに応用した例を以下に示す．

2．意味の空間的表現

意味の空間とは．何となく近いと思う概念どうしが近くに，何となく遠いと思う概念どうしが遠くになるように配置されたユークリッド空間である．近い，遠いの判断は，ユーザが論文を検索する時に，関係ある概念として指定するかどうかにより，なされる．「何となく」という点が重要であって，ユーザは，上位下位関係，部分全体関係などの論理的関係を使って，自分の考えを分析しながキーワードを与えることは苦手だが，見せられた空間の中から関係あるものをピックアップすることは容易にできる．

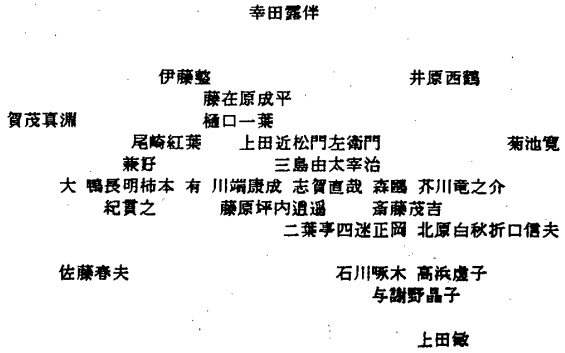

図1　システムの作成した初期空間の例

空間中の概念は，選ばれた論文のタイトル中の単語を付け加えることにより増加していく．文学の論文検索の場合，初期状態としては，作品名と作家名が与えられている．原理的には空間の次元や軸は，あらかじめ与えられておらず，空間

Representation and Learning of the Space of Wordmeanings for Information Retrieval Systems

Koichi HORI, Seinosuke TODA, Hisashi YASUNAGA
National Institute of Japanese Literature

構成の結果として定まるが，実装したシステムでは3次元に押さえている．論文タイトル中の単語の共出現関係を手がかりとして，システムが作成した初期空間の例を図1に示す．この空間が何かを意味することは，見る人が見ればわかる．すなわち，意味の空間的表現は，意味の直接的表現というより，ユーザが頭の中で抱いている意味を引き出すための媒介的な役割を果たしている．

3．意味空間の学習

システムの構成を図2に示す．学習の過程については，文献〔1〕〔2〕にゆずることとして，本稿では，単語の意味の空間的表現とその学習が，論文検索のためにどのような効果をもたらすかについて，実例を示すことにする．

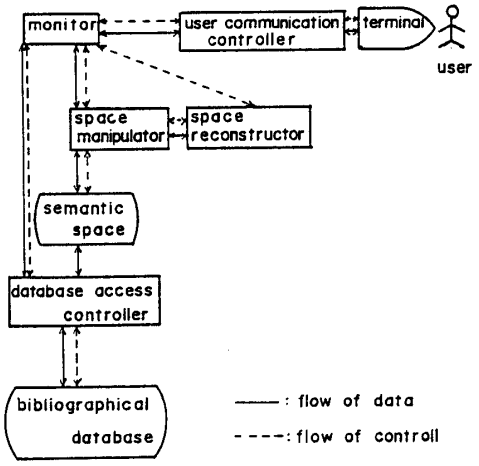

図2 意味空間学習システムの構成

ある文学者は，「日本文学と宗教の関係」について検索を試みた．従来のキーワードによるシステムだと，「日本文学」と「宗教」をキーワードにすると，当然ながら，それらをキーワードとしてもつ論文しか出力されない．ところが，筆者らのシステムの場合は，宗教の近くに「仏教文学」と「キリシタン文学」という概念が配置されているので，それらも関係あると指示すれば，それらの論文も出てくる．従来のシステムでも，それらをキーワードとして与えればいいわけだが，ユーザはなかなかうまいキーワードを思い付かないものである．それに対して，空間にして見せられれば，その中から関係あるものを指摘するのは非常に容易である．また，論文を出力する際も，タイトルの似たものが近くになるように自動的に配置されるので，大量に出力しても，欲しいものを簡単に見付けることができる．宗教と文学といえばキリシタン文学と思うユーザ用の空間では宗教とキリシタン文学が近付き，仏教文学だと思うユーザ用の空間では宗教と仏教文学が近付くことになる．さらに，選ばれた論文のタイトルから関係ある概念が空間に追加されてゆく．

シソーラスとの違いは，任意の関係を空間にとりこめることで，通常のシソーラスでは，「宗教」の下位概念は「キリスト教」ということになるが，日本文学という文脈のもとでは「キリシタン文学」という概念を考えなければならない．従来の研究ではそのようなさまざまな関係をあらかじめ分析して並べあげておこうとしたのに対して，筆者らのシステムでは「何となく」関係があるとして，空間配置という形に表現されることになる．そのうちどれとどれはどういう関係にまとめあげられるとわかったならば，そこを従来の意味表現に接続すればよいが，その組織だった方法については，今後の研究課題である．

4．むすび

単語の意味の空間的表現法を考え，その学習メカニズムを与えた．その有用性が，論文検索システムという実用場面で実証された．何となく近いものが近くにある空間という説明は，ユーザに容易に理解してもらえ，概して評判が良い．ユーザとの親和性という面でも本稿で提案した方法は優れている．

参考文献
〔1〕Hori, K. et al.: Learning the Space of Wordmeanings for Information Retrieval Systems, Proc. COLING86, Bonn(1986).
〔2〕堀：単語の意味の学習について，コンピュータソフトウエア（日本ソフトウェア科学会誌）（掲載予定）．