あいまい情報検索

岩井 壮介・中村 清彦

1. まえがき

情報検索は，人間とコンピュータの記憶する知識・データベースとの対話のプロセスである．情報検索者の目的とするものは，新しい研究や設計，複雑な問題に対する意思決定のための総合的知識・新しい概念を獲得することにある．しかし，現在のデータベースにおける知識蓄積の方法は，たとえば文献データベースにみられるように，1つの文献について題目，著者名，内容の要約であるいくつかのキーワード，あるいはアブストラクトが組として記憶されており，各文献の表わす概念を総合するための情報を含んでいない．したがって，コンピュータの提供するものは知識の断片であって，総合的知識・新概念の生成は，まったく検索者の創造的能力にたよっている．この現在の情報検索の限界を乗り越えるためには，人間とデータベースの間に，知的なインタフェイスとしてのコンピュータ利用技術を導入し，人間とコンピュータとの接点を，知的により高度なレベルへ引き上げることが必要であろう．

人間は対話における相手の短い表現，接触する外界でのちょっとした出来事から，それらが示す事柄以上の情報を類推によって得る能力をもっている．また，頭の中のあいまいな部分を積極的に質問することによって，相手のもつ概念や外界の出来事に対する的確な認識をもつにいたる．ここで紹介する内容は，人間のもつ，このような類推機能，概念学習の能力を知識総合のためのコンピュータ推論機構として実現し，それを情報検索システムに応用する試みに関するものである[1],[2].

いわい そうすけ，なかむら きよひこ
京都大学 工学部

2. 類推および概念学習過程のFuzzy集合によるコンピュータ・アルゴリズム化

総合的知識・新しい概念は，多くの既存の概念を関係づけ，総合することにより生まれる．たとえば，文献検索において検索者の要求する文献は，新しい概念を創造するのに用いられる既存の概念と考えることができる．

2.1 知識空間——知識および既存概念のコンピュータ内部表現

知識のコンピュータ内部表現として，M. R. Quillianの提唱したSemantic Memoryがある[3]～[7]．Semantic Memoryは，知識のグラフ構造化表現であって，グラフのノードおよびリンクは，それぞれ概念単位，および単位間の関係を表わす．このような知識表現を用いるとき，概念単位の集合として表わされる1つの概念は，概念単位間のリンクを介して他の概念（概念単位の集合）に関連づけられることがわかる．概念単位間の関係は，何らかの意味での単位間の親近性にもとづいてきめられるものであって，その定量化は，対象とするデータの性質に依存する．いま図

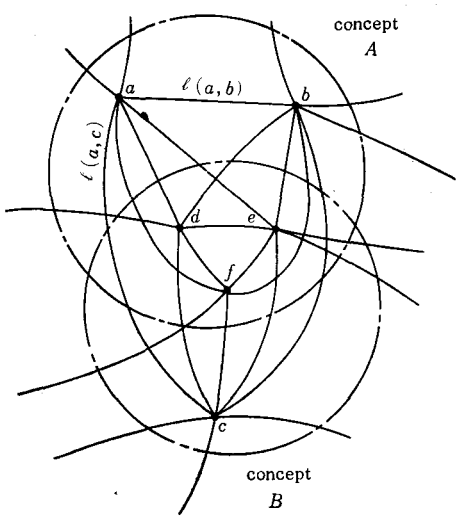

図 1 先験知識空間

に示すように，1つの概念に共通に含まれる概念単位対 (a,b) は，そうでないもの，たとえば (a,c) より概念形成のうえから親近性が大きく，同図 (d,e)，(e,f) あるいは (d,f) のように多くの概念に共通に用いられるものほど，その度合が大きくなると考えることができる．したがって，概念単位対 (x_i,x_j) の同一概念同時出現の相対頻度値 $l^{-1}(x_i,x_j)$ を用いて親近性の度合を定量化することができる*．これは，心理学におけるSimilarity Measureに対応するものである[9]～[11]．この値の逆数 $l(x_i,x_j)$，すなわち概念単位間の抽象的な距離を，リンクの重みとして与えることにより得られる重みつきグラフを知識空間 X とよぶことにする．知識空間 X は，コンピュータ・メモリーに与えられた先験知識であり，その中で，既存の概念は X のノード，すなわち概念単位の部分集合として表わされている．コンピュータは，この先験知識を用いることにより，おそらくはいくつもの既存の概念にわたると思われる情報検索者の求める新概念を，既成概念間のリンクをたどって大局的に学習把握することができる．要求概念のコンピュータによる大局的把握があってはじめて，コンピュータは適切かつ詳細な情報を提供することが可能となり，人間とデータベースの間の知的インタフェイスとしての機能を果すことができよう．

2.2 Fuzzy化関数による類推機能の定量化と要求概念のコンピュータ学習過程

いま，コンピュータに，ある1つの概念単位 x_s が検索者の要求概念に含まれることを知らされたとする．コンピュータは，この明確な情報とともに，他の概念単位 $x(\in X)$ についても，x_s からの距離 $l(x_s,x)$ にもとづいて，検索者の要求概念に対する帰属の度合を類推*することができる．この度合を定量的に与えるものとして，図2(a)に示すようなFuzzy化関数 $f_{x_s}(x)$ が考えられる．すなわち，コンピュータの先験知識空間 X に含まれる概念単位 x には $f_{x_s}(x)$ というメンバシップ値が与えられ，X 上に，図示のようなFuzzy集合 F_{x_s} が生成される．F_{x_s} は，x_s が未知要求概念に含まれるという後験情報，およびそれからの類推によってコンピュータが得た付加的情報と考えることができ，これはまた，検索者の要求概念に対するコンピュータの最初の認識 $\tilde{C}_0=F_{x_s}$ と考えることができる．ついで，概念単位 x_q がやはり要求概念に含まれるという情報をコンピュータが得たとする．このときコンピュータの認識は，同図(b)に示すように $\tilde{C}_1=\tilde{C}_0\cup F_q$ に変更される．また，x_q が要求概念に含まれないという情報を得たときには，同図(c)に示すようなFuzzy化関数 $\bar{f}_{x_q}(x)$ $(=1-f_{x_q}(x))$ を用いることにより，x_q 以外の単位 x の要求概念への帰属の度合の減少を定量的に

---

* L.B. Doyleは，キーワードを概念単位に，キーワード集合で表わされる文献を個々の既存概念と考えることにより，キーワード対の同一文献同時出現の頻度値を用いて，キーワード間の親近性を与えている[8]．

* ここで用いた言葉“類推”は，心理学における“類推(analogical reasoning)”とは異なり，むしろ“汎化(generalization)”や転移効果“(transfer effect)”に対応する．R.C. AthinsonおよびW.K. Estesは，similarityにもとづく汎化の集合論的モデルを提唱している[12]．

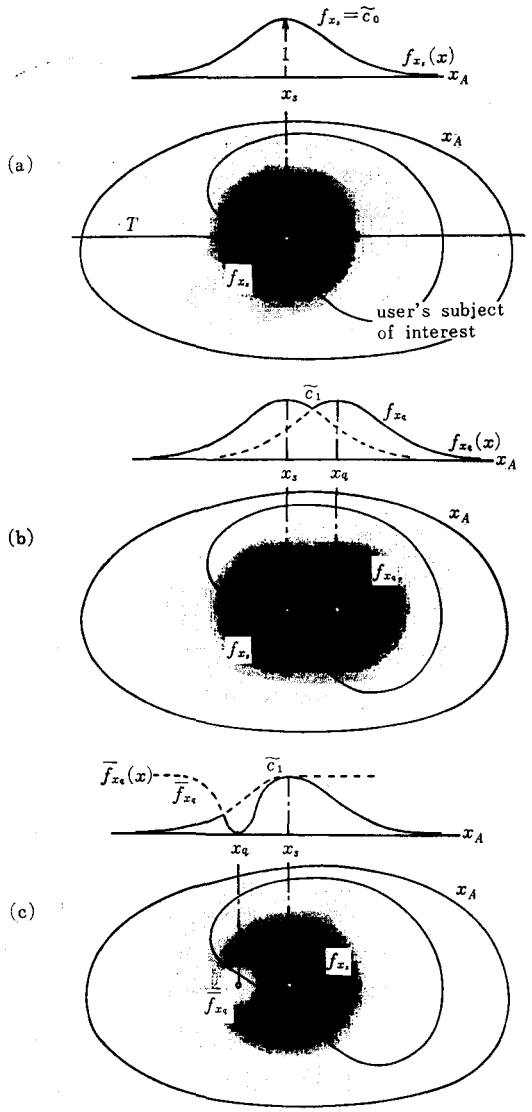

図 2 Fuzzy 化関数 $f(x)$および$\bar{f}(x)$による類推機能の定量化とコンピュータの認識概念の Fuzzy 集合 $\tilde{C}$による表現

与えることができる．$\bar{f}_{x_q}(x)$をメンバシップ関数とする Fuzzy 集合 $\bar{F}_{x_q}$は，やはり x_qが未知概念に含まれないという後験情報，およびそれからの類推によってコンピュータが得た付加情報と考えることができる．このとき，コンピュータの認識は，同図に示すように，$\neg\tilde{C}_1=(\neg\tilde{C}_0)\cup(-\bar{F}_{x_q})$，すなわち $\tilde{C}_1=\tilde{C}_0\cap\bar{F}_{x_q}$と変更される．

このように，コンピュータの先験知識の空間Xに存在する概念単位が検索者の要求概念に含まれ

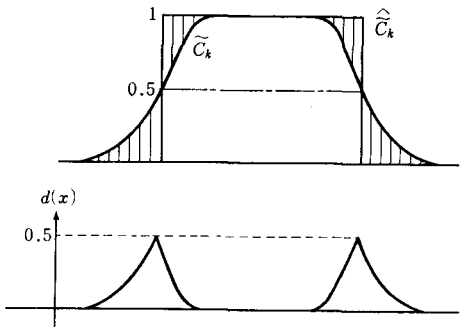

図 3 A. Kaufmann による Fuzzy 集合のあいまいさの指標

るかどうかという後験知識獲得のプロセス——Q-Aプロセス——の進行とともに，コンピュータの認識は$\tilde{C}_0,\tilde{C}_1,\tilde{C}_2,\cdots$と学習更新されることになる．この場合，情報検索者の最初の指示x_sによって生成される $\tilde{C}_0$ を除いて，以降における認識の更新は，要求概念に対するその時点での認識にもとづいて，コンピュータ自身が質問項目としての概念単位を次にのべる方法で選択することにより遂行される．人間は，1でのべたように，自分のもつあいまいな認識をできるだけはやくはっきりしたものにするよう質問を行なう．この能力をアルゴリズムとしてコンピュータに与えるには，質問概念単位として，それに対する検索者からの回答によって，要求概念に対する認識のあいまいさが最も減少するようなものを選択させるようにすればよい．上述のように，コンピュータの識識は Fuzzy 集合として表わされている．第k回目のQ-A を終わった時点での認識を $\tilde{C}_k$ とする．Fuzzy 集合 $\tilde{C}_k$のあいまいさの定量化として，A. Kaufmann の与えた Index of Fuzziness $I(k)$ がある[13]．$I(k)$は，図3にように，$\tilde{C}_k$ と自乗誤差最少の意味において最も近い通常集合 $\tilde{\tilde{C}}_k$ を考えるとき，その自乗誤差の値に相当する．なお，同図に示す $d(x)$は，概念単位xの $\tilde{C}_k$ への帰属・不帰属のあいまいさを与える尺度と考えることができる．

さて，選ばれた質問概念単位に対する検索者の回答，Yes（要求概念に含まれる），No（含まれ

ない）はまったくわからない．したがって，コンピュータは，それぞれの場合に新しく生成される認識（Fuzzy 集合）の I 値の平均値が最少となるようなものを知識空間の中から選ぶ．すなわち，第 k 回目の質問概念単位 x_k としては，$\tilde{C}_{k-1} \cup F_{x_k}$ および $\tilde{C}_{k-1} \cap \bar{F}_{x_k}$ の I 値の平均が最も小さいものを検索者に提示する*．この値が小さくなるのは，大まかに言って，先述の帰属・不帰属のあいまいさ $d(x)$ が最大(0.5)となる附近，すなわち，Fuzzy 集合 $\tilde{C}_{k-1}$ の境界の附近である．

以上の説明において，検索者はコンピュータからの質問，“x は貴方の考えている内容に含まれるか”，に Yes, No で明確に答えられるものとしている．すなわち，図4に示すように，検索者は自分の要求概念を，通常集合 C として自分の頭の中にもっているとしている．これは次のような考えにもとづくものである．検索者は，C を構成するすべての概念単位を洩れなく表明することはできない．このことは，検索者が何らかの創造的な仕事，たとえば研究，設計などに着手し，その発想を頭の中に描き出したとしても，それが着想の段階であるかぎり，それを詳細にわたるいわゆる procedure として記述できないことと似ている．しかし，procedure 全体を陽には指示できないが，その着想に関連のある概念単位とそうでないものとは，それが質問されたとき比較的容易に判断できるであろう．

2.3 学習の終了と要求情報の出力

コンピュータの知識空間 X に含まれる概念単位が有限の場合，すべての単位について質問を行なうことにより，コンピュータは要求概念 C を完全に知る，すなわち認識概念を表わす Fuzzy 集合 $\tilde{C}_k$ を通常集合 C に一致させることができる．しかし，実際問題としては，ある程度まで学習認識した時点で，換言すれば認識概念のあいまいさ I 値がある値以下になったところで Q-A プロセスを打ち切ることになる．この場合，次のような問題が生ずる．**要求概念 C は，図4に示すような連結的な集合とは限らない．**要求概念が漠然として広いとき，C は非連結な通常集合の集まりと考えるのが自然な場合がある．このようなとき，$\tilde{C}_k$ が C の一部分に近づいたとき学習が終了することがおこりうる．$\tilde{C}_k$ と C の不一致の度合を示す尺度として，$I(k)$ の場合と同様，両者の間のユークリッド距離 $D(k)$ を定義することができる．

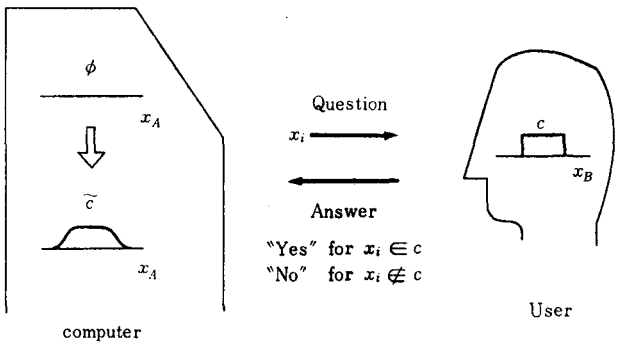

図 4　コンピュータと情報検索者の対話の模様

2.1にのべたように，検索者の要求する知識，既存概念は，図5に示すように知識空間 X における概念単位の通常集合 D_s として表わされている．図に示す R_s は，D_s に含まれる概念単位が同概念を表わすうえにおいて同等の重みをもつと仮定したときの，D_s の内容のうち学習認識された要求概念 $\tilde{C}$ に含まれる部分の割合を意味する．コンピュータは，検索者により指定された以上の R_s 値をもつ概念単位を，要求知識として出力する．

3. 文献情報検索システムへの応用

以上の考え方は，706頁の脚注にのべたようにキーワードを概念単位に，キーワードの集合を既存概念に対応させることにより，文献検索システムに応用することができ，その場合，コンピュータは，文献検索者が頭に描いている領域に属すると思われる文献名を，大局的に判断して提供する

---

* この場合，$x_k \in X$ のすべてについて I 値の平均を求めるのではなく，すでに用いられた概念単位以外のもののうち，$\tilde{C}_{k-1}$ のメンバシップ値がある値以上のものについて探索する．

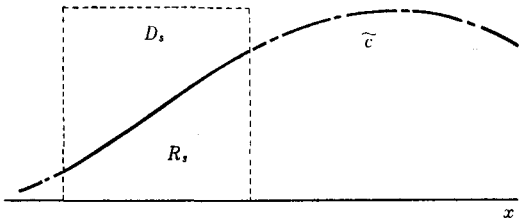

図 5 既存概念 D_s の内容のうちコンピュータの認識 $\tilde{C}$ に含まれる部分の割合 R_s

ことになる．また，このシステムを従来の文献データベースと結合することにより，要求概念にマッチした詳細な情報を提供することが可能になる．

試作したシステムは，機械工学に関する技術論文に含まれる812個のキーワード，5085組のキーワードと文献の対から求めた知識空間 X をもち，検索者との対話は，電話回線を介しての TSS で実行される．類推機能を定量化する Fuzzy 化関数としては，$f_{x_i}(x)=\exp(-\alpha|x-x_i|^2)$ を用いており，図6に示すように，パラメータ α によって類推のおよぶ範囲を適宜規定することができる．図7は対話実行例である．

[PROC. 1] は，検索者による α の初期指定である．α 値は，Q-A プロセスを通して一定に保たれるのではなく，コンピュータによって適応的に変更される．たとえば，図6において，x_j を新質問キーワード，x_i をその近傍に存在する過去に質問に用いられたキーワードで，x_j に対する検索者の回答が，x_i に対するものと相反していたとする．このとき，要求概念の境界が x_j の近傍に存在することが想像される．このようなとき，α 値を大きく，すなわち $f_{x_i}(x)$ の形をシャープにして，あまり類推に頼らない緻密な学習を行なう．

[PROC. 2] は，検索者による最初のキーワード x_s の提示である．ここでは仮に，検索者の要求する概念 C は，機械加工の自動化を内容とする1つの文献，"Machining Centre Does 1500 Operations"（文献番号 1447，キーワード $C=$ {MACHINING CENTRE, NUMERICAL CONTROL, CUTTING, CRANKSHAFT}）に一致すると仮定している．そこで検索者は，x_s として上記概念内容を代表する MACHINING CENTRE を入力している．コンピュータは，最初に得た認識 $\tilde{C}_0$ において，0.1以上のメンバシップ値を有する12個のキーワードを表示している．

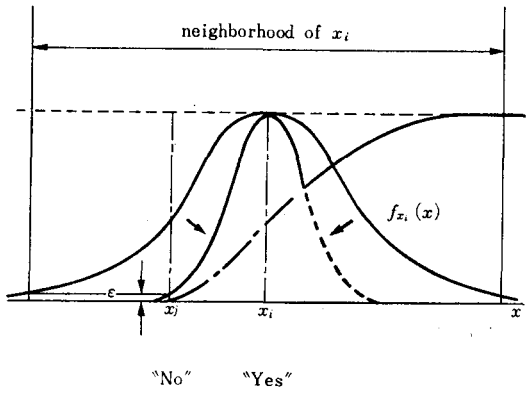

図 6 類推機能定量化のための Fuzzy 化関数 $f_{x_i}(x)=\exp(-\alpha|x-x_i|^2)$ と α の変更による類推範囲の適応的変化

[PROC. 3] は，コンピュータによる第1回目（$k=1$）の質問，それに対する検索者の回答を示す．2.2にのべた選択基準により，RECIPROCATING COMPRESOR が質問キーワード x_1 として選ばれ，検索者は $x_1 \notin C$ であるから No を回答している．その結果，コンピュータの認識は，$\tilde{C}_1=\tilde{C}_0 \cap \bar{F}_{x_1}$ に変化し，[PROC. 2] で0.61であった x_1 のメンバシップ値は 0(<0.1)となって表示から削除されている．また，$\tilde{C}_1$ のあいまいさの指標 $I(1)$ は 0.057 となり，$I(0)$ より減少している．この実行例では，I 値がそれまでの最大値の半分以下に減少した時点で学習終了としている．$k=11$，すなわち，第11回目の質問項目MILLING に対して No が回答された時点でこの条件が満足されている．このとき，表示されたキーワードは，メンバシップ値が1に近いものと0.1に近いものに二分されており，学習された概念 $\tilde{C}_{11}$ が十分通常集合に接近していることがわかる．

[PROC. 4] は，検索者の文献出力要求に対す

```
[PROC.1] Q: ALPHA ?
         A: 0.006

[PROC.2] Q: WHAT KIND OF INFORMATION DO YOU NEED ?
           PLEASE INPUT A KEYWORD OF THE INFORMATION.
         A: MACHINING CENTRE

                I(0)=0.06572
                  <KEYWORD>      <MEMBERSHIP FUNCTION>
                 RECIPROCATING COMPRE0.61
                 CONTACT RATIO        0.10
                 MACHINING            0.14
                 CRANK SHAFT          0.74
                 MACHINE TOOL         0.56
                 LUBRICATION          0.45
                 NUMERICAL CONTROL    0.88
                 CUTTING              0.21
                 CHATTERING           0.11
                 SURFACE ROUGHNESS    0.25
                 MILLING              0.68
                 MACHINING CENTRE     1.00

[PROC.3] (1)Q: IS RECIPROCATING COMPRESSOR
             CONTAINED IN THE INFORMATION ?
           A: NO
                I(1)=0.05753
                 CONTACT RATIO        0.10
                 MACHINING            0.14
                 CRANK SHAFT          0.12
                 MACHINE TOOL         0.56
                 LUBRICATION          0.45
                 NUMERICAL CONTROL    0.88
                 CUTTING              0.21
                 CHATTERING           0.11
                 SURFACE ROUGHNESS    0.25
                 MILLING              0.68
                 MACHINING CENTRE     1.00

         (2)Q: IS LUBRICATION
               CONTAINED IN THE INFORMATION ?
           A: NO
                I(2)=0.04811

         (3)Q: IS NUMERICAL CONTROL
               CONTAINED IN THE INFORMATION ?
           A: YES
                I(3)=0.04275

          ------------------------

         (11)Q: IS MILLING
                CONTAINED IN THE INFORMATION ?
            A: NO
                I(11)=0.03177
                  RECIPROCATING COMPRE0.10
                  CRANK SHAFT          1.00
                  AUTOMATIC LATHE      0.64
                  NUMERICAL CONTROL    1.00
                  NC LATHE             0.13
                  CUTTING              1.00
                  THERMAL DEFORMATION  0.17
                  CHATTERING           0.11
                  MACHINING CENTRE     1.00

[PROC.4] (1)Q: HOW MUCH AMOUNT OF THE INFORMATION
               SHOULD BE CONTAINED IN THE DOCUMENTS ?
                PLEASE INPUT THE AMOUNT AS PROPORTION.
           A: 0.9

[OUTPUT]     NUMBERS OF THE DOCUMENTS YOU NEED
              1416
              1447
              1448
                -- END OF DATA

         (2)Q: HOW MUCH AMOUNT OF THE INFORMATION
               SHOULD BE CONTAINED IN THE DOCUMENTS ?
                PLEASE INPUT THE AMOUNT AS PROPORTION.
           A: 0.5

              NUMBERS OF THE DOCUMENTS YOU NEED
               1302      1396
               1339      1403
               1341      1406
               1342      1416
               1344      1417
               1347      1447
               1353      1448
               1392        -- END OF DATA
```

図 7 試作文献検索システムの対話実行例

るコンピュータの応答を示す．検索者が図5の意味において，$\tilde{C}_{11}$ と90%以上の概念内容上の重なり($R_s=0.9$)を有する文献を要求したときは，コンピュータは1447を含む3つの文献（1447，1416＝“Two Typical Applications of CN”，1448＝“The Economical Efficiency of Keyboad Programing NC-machine”）を出力している．概念上の重なりが50%以上であればよいという指定に対しては，コンピュータは15の文献を出力している．この場合，たとえばその中の1つ，1406＝“Manufacture of Hollow Rivets with Automatic Lathes”は，自動旋盤を用いた特殊な加工事例についての文献であり，要求内容に直接結びつくとは言えないものも含まれてくる．

ここでのべたシステムの学習効率は，検索者の求める概念Cの大きさに対し，類推のおよぶ範囲を規定するパラメータαの値をどのように選ぶかに依存する．Cが漠然として広い（Cを代表するキーワードが多く，したがってそれらが，知識空間における多くの文献に関連している）場合，αを小さくして類推のおよぶ範囲を拡げるのが効果的であり，Cが限定されている場合には，αを大きく，Fuzzy 化関数の形をシャープにして，あいまいな情報をとり込まないほうが賢明であると思われる．事実，試作システムを用いての実験では，Cの大きさ*によって，学習効率**を最大にするαの値が存在することがわかっており，またそのときの効率自身の値は，要求概念C（キーワード集合）の知識空間における分布の形に依存している．すなわち，キーワード集合が，空間のある領域に凝集しているとき，その効率は大きくなり，空間に不均一に分散しているとき小さくなる．

4. むすび

本稿は，人間の有するすぐれた情報処理能力，特に類推機能と概念学習能力を，コンピュータに付与する試みについてのべたものである．ここで

---

* ここでは，Cの含むキーワード数をとっている．

** 学習終了時のQ-A回数k，そのときの$\tilde{C}_k$とCとのユークリッド距離$D(k)$を直交座標系にとったとき，点$(k, D(k))$が原点に近いほど学習効率が大きいと考えている．

はふれなかったが，われわれのシステムにおいて，類推機能および概念学習能力を発揮するうえに重要な役割をもつコンピュータの先験知識空間は，コンピュータが検索者との対話により新概念を修得することによって学習変更されるようになっている．さらに，たとえば，概念単位を概念説明上のレベル，あるいは属性の重要度によって区別し，知識空間に階層構造を導入するなど，知識表現の高度化，Q-Aの各ステップの改良などにより，システムの能力を向上させることが可能と思われる．

1980年代におけるシステム科学の重要な課題の1つは，多様化時代における各種複雑な問題に対し，人間の有する創造的能力が十分に活用されるようなシステムの設計であろう．1人1人の人間の理解しうる知識の範囲は限られている．多くの専門分野の知識を総合してはじめて可能となるこれらの問題の解決には，膨大な情報との対話が必要である．それには，今日，すさまじい勢いで蓄積の続いているコンピュータ記憶：データベースと人間との接点に，新しいコンピュータ利用技術を導入することによって，対話のレベルを引き上げ，コンピュータに知的作業を分担させることが不可欠であろう．

参考文献

[1] 中村，岩井：Fuzzy 集合の要素間関係に対する位相的構造の導入と人間の類推学習行動のモデル化，システムと制御，Vol. 24, No. 4, pp. 52～60 (1980)

[2] K. Nakamura & S. Iwai : Topological Fuzzy Sets as a Quantitative Description of Analogical Inference and its Application to Question-answering System for Information Retrieval; IEEE Trans. on System, Man and Cybernetics (to appear)

[3] M. R. Quillian : Semantic Memory; Ph. D. dissertation, Carnegie Institute of Technology (1966)

[4] M. R. Quillian : Semantic Memory ; in Semantic Information Processing (M. Minsky ed.), M. I. T. Press, pp. 227-270 (1968)

[5] J. R. Carbonell & A. M. Collins : Natural Semantics in Artificial Intelligence, Proc. of the Third International Joint Conference on Artificial Intelligence, pp. 344-351 (1970)

[6] J. R. Carbonell : AI in CAI : An Artificial-intelligence Approach to computer-assisted Instruction, *IEEE Trans. on MMS-11*, No. 4, Dec., pp. 190-202 (1970)

[7] R. Fugmann, H. Nickelsen I. Nickelsen & J. H. Winter : Representation of concept Relations Using the TOSAR System of IDC: Treatise III on Information Retrieval Theory, *J. of the American Society for Information Science*, Sept.-Oct., pp. 287-307 (1974)

[8] L. B. Doyle : Indexing and Abstracting by Association; *American Documentation*, Vol. 13, pp. 378-390 (1962)

[9] L. Sjöberg : A Cognitive Theory of Similarity ; Göteborg Psychological Reports, No. 10 (1972)

[10] R. A. M. Gregson : Psychometrics of Similarity; New York : Academic Press (1975)

[11] A. Tversky : Features of Similarity ; *Psychological Review*, Vol. 84, No. 4, pp. 327-352 (1977)

[12] R. C. Atkinson & W. K. Estes : Stimulus Sampling Theory; In R. D. Luce, R. R. Bush, and E. Galanter (Eds.) Handbook of Mathematical Psychology, Vol. II, New York : Wiley, pp. 121-268 (1963)

[13] A. Kaufmann : An Introduction to the Theory of Fuzzy Subsets; Vol. 1, New York : Academic Press (1975)