知能と複雑系 119－17
（2000. 1. 13）

概念ベースにおける属性数の検討と概念間の関連度計算方式

入江 毅　　渡部 広一　　河岡 司
同志社大学大学院 工学研究科 知識工学専攻
〒610-0394 京都府京田辺市多々羅都谷 1-3

あらまし 柔軟な利用者インタフェースを持つ情報処理システムの中核となる機構は概念ベースと概念の関連度に関係する連想機能と考えられる．自動学習や自動精錬の観点から概念ベースは可能な限り単純な構造でなければならない．このため，本稿では，概念ベースにおいて各概念が持つ属性数を固定数 N とすることとし，より適切な N について実験により評価した．さらに，概念の 2 次属性の一致数までを評価の対象とし，これを非線形に変調する関連度計算方式が人間の感覚により適合することを実験により示した．

キーワード 概念ベース，関連度，連想機能，概念連鎖，固定属性数，変調関連度

An Evaluation of the Number of Attributes in Concept-Base and Measuring the Degree of Association between Concepts

Takeshi Irie , Hirokazu Watabe, Tsukasa Kawaoka
Department of Knowledge Engineering and Computer Science, Faculty of Engineering,
Doshisha University, Tanabe, Kyoto 610-0394

Abstract It is thought that the main elements of information processing systems which have friendly user interfaces are the concept-base and the association mechanism based on the degree of association between concepts. It is expected that a structure of the concept-base is as simple as possible since the concept-base has to be expanded and refined automatically by automated learning. In this paper, it is assumed that each concept in the concept-base has fixed number of attributes N, and the proper fixed number N is derived by experiments. Furthermore, the new method measuring the degree of association between concepts is proposed. This proposed method, which modulates the matching number of second stage attributes by nonlinear function, can output the closer degree of association to that decided by human judge.

key words concept-base, a degree of association, association mechanism, chain concept, fixed number of attributes, modulated method

1. はじめに

情報処理システムは人間社会の様々な分野で活用され，もはや欠かすことのできない要素となっているが，これまでの発展のほとんどは，機能面，性能面での高度化に起因するものであり，知的な観点での著しい進展はあまり見られない．しかし，今後，情報処理システムの機能・性能面での高度化が進めば進むほど，人間社会の知的な情報処理への憧れはいっそう強まってくるものと思われる．そこで，我々は「知的」の本質の解明というよりは，知的と呼んでもあまり違和感がなく，しかも，従来とは異なるメカニズムで，情報処理の高度化につながるような現実的なメカニズムの創出を目的とした，「知的判断メカニズム」の実現に挑戦している．

知的判断メカニズムを実現するためにはその中核機構として，様々な判断を汎用的に行うための知識資源である概念ベースと，概念間関連度計算方式を利用した連想機能を備える必要があり，この連想機能を利用することによって人間らしい，様々な判断を可能とする．ここで，概念ベースは維持管理の面から可能な限り単純な構造であることが望まれる．この単純化された構造を持つ概念ベースを連想概念ベースと呼ぶ．

本稿では，連想概念ベースの構造を決定する要素として，各概念の属性数を固定の N 個とし，より適切な N について，概念間関連度計算方式により知的判断の妥当性の観点から評価した．さらに，2 次属性一致個数に変調を行うことによって，概念連鎖による概念の関連性をより適切に表現する，新たな概念間関連度計算方式について述べている．

以下，2 章で，まず概念間関連度計算方式の概要を，3 章で我々のねらいとする知的判断メカニズムに必要となる連想概念ベースの構造について述べ，4 章で概念間関連度計算方式による連想概念ベースの構造の評価を述べている．さらに，5 章以下で従来方式とは異なる，新たな概念間関連度計算方式の提案とその評価について述べる．

2. 概念間関連度計算方式

概念間関連度（以降単に関連度と呼ぶ）とは，概念の関連性を定量的に評価するものである．ここで，関連性は絶対的な尺度ではなく，相対比較によって定義する．例えば，「自動車」と「乗り物」の関連性と「自動車」と「花」の関連性を考慮した場合，人間ならば「自動車」と「乗り物」の方が関連が深いと容易に判断するが，この関連性を関連度（図1）と呼ぶ数値で表現し，その大小関係を考慮することにより，コンピュータ上で実現するものである．このような定義は厳密ではないが，人間の会話における意図理解が現実的に，この程度で行われていることから，知的判断メカニズムの基盤としては妥当と考える．

Rel（自動車，車）=0.87
Rel（自動車，乗り物）= 0.71
Rel（自動車，花）= 0.01
Rel（A，B）：概念Aと概念Bの関連度

図1：関連度の例

具体的には，概念間の暗黙の論理関係を積極的に活用し，概念の属性集合の関連性を評価するために，概念連鎖により概念をn次属性レベルまで展開し，概念間の論理関係が薄れたところで記号としての属性一致数を評価することにより算出する考え方であるが，今回は2次属性まで展開する方式[1]を用いた．

3. 連想概念ベース

3.1 概念構造

基本概念ベースは，複数の国語辞書等の語義文から自立語の出現頻度に基づいて属性とその重みを獲得し，更に，それの自己参照による新たな属性の追加，及び不要な属性の統計的な除去からなる精錬を行うことによって，完全に機械構築した約4万語の概念により構成されており，概念Xを属性語のa_iとその出現頻度等を考慮した重みw_iの集合で構成し，概念属性数（以降単に属性数と呼ぶ）は各概念によって異なる[2]．しかし，連想概念ベースの構造は，重み情報が付加されていない属性語a_iの集合であり，ある概念に対する属性数は全ての概念において等しいN個であるとする．

これは，自動学習（概念の自動的追加）と機械精錬の容易性の面からできるだけ単純な構造にすることが必要となるためであり，重み情報を付加させると，ある概念に新たな適切属性の追加や不適切属性の除去が行われ

た場合，その都度，概念に属する全ての属性の重み情報を精錬する必要性が生じるが，これらの処理を機械的に行うことは極めて困難である．また，概念属性数を可変とした場合，新たに出現した属性を概念との関連度等により評価を行い，取得するか否かの判断を行う際，その絶対的な閾値を決定する必要が生じる．しかし，現状の関連度計算方式においては，相対的な尺度を用いることを前提としており，算出される関連度にかなりばらつきが生じる．例えば同義語（語形は異なるが、内容が（ほとんど） 同じ関係にある語）の関連度を算出した場合においても，その値は個々においてばらつきが生じる．よって，関連度の閾値の決定は困難であるため，現時点での連想概念ベースにおいては属性数を固定とした．

連想概念ベースの概念構造

概念X:{ $a_1, a_2, \cdots, a_N$ }

3.2 初期連想概念ベース

連想概念ベースは基本概念ベースを原点として構成されるが，初期の段階では適切な固定属性数Nが決定されない．そこで，初期の連想概念ベースは固定属性数Nの初期値を100として基本概念ベースを再構成した．

初期値を100とした理由は基本概念ベースの属性数の分布（図2）において，属性数が100以上のものが4%以下であり，Nの初期値は100で十分であると判断したためである．この再構成法のアルゴリズムは，属性が100個に満たない場合は，属性の属性，つまり，概念連鎖を再帰的に用いることにより属性を取得するものであり，100個以上である場合は，重みの101番以上を切り捨てた．ここで，新たに取得された属性の概念に対する適切さの厳密な評価は行わず，再構成された初期連想概念ベース（以降単に概念ベースと呼ぶ）に属する属性語は，概念をそれなりに適切に表現するものと仮定する．

概念ベースの概念構造

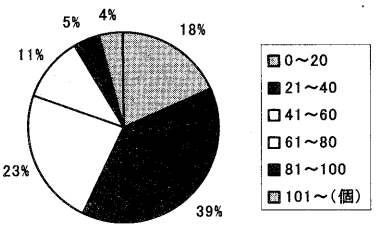
概念X:{ a_1, a_2, $\cdots$, a_{100} }

4. 固定属性数Nの評価

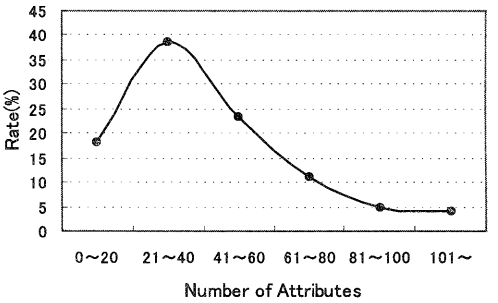
4.1 概念属性の関連度による順位付け

概念ベースにおいて，より適切な固定属性数Nを関連度を用いて実験により評価する際，100個の属性の選択方法は無数にある．そこで，仮の固定属性数NをN$_T$=40として概念とその属性100個との関連度を算出し，その関連度を重み情報として属性を関連度順に並べた．ここで，仮の固定属性数N$_T$は，基本概念ベースの全概念属性数の平均値を1の位で四捨五入したものである．

なお，実験により，N$_T$は20, 30, 40, 50で殆ど後の論理に影響しないことを確認している．



(a)



(b)

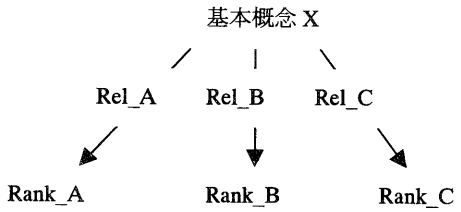
図 2:属性数の分布

4.2 評価尺度の作成

ある基本概念Xに対して，図3のように関連が深い概念:Rank_A，関連がある概念:Rank_B，関連がない概念:Rank_Cを想定し，これら4つの概念を一組とした評価尺度を人手によって500組作成した．評価尺度の例を表1に示す．

表1:評価尺度の例

基本概念X	Rank_A	Rank_B	Rank_C
草	雑草	植物	車
火	炎	火事	海
話	物語	本	雲
木	材木	家	道路
円形	丸い	四角	音楽



Rel_Y ： 基本概念 X と Rank_Y との関連度

図 3:評価尺度

4.3 評価実験方法

4.2で作成した評価尺度からランダムに100組選択し，10通りの属性数N（N=10,20,30, …,100）によって100組のRel_Yを算出する．算出されたRel_Yを次の評価式(1)(2)(3)[3]により評価する．

$$F_1 = \frac{R_a - R_c}{\sigma_a + \sigma_c} \quad \cdots \quad (1)$$

$$F_2 = \frac{1}{1+wg} \quad \cdots \quad (2)$$

$$F_d = F_1 \times F_2 \quad \cdots \quad (3)$$

ここでF_1は基本概念に対して，関連が深いものとあまり関連がないものとの距離を評価するための指数であり，R_a, R_cは図3におけるRel_A, Rel_Cの全サンプルに対する平均値であり，σ_a, σ_cはおなじくRel_A, Rel_Cの標準偏差である．このF_1が大きくなるほど距離が大きい，つまりより良い結果であることを表す．また，F_2はF_1のように大局的な関連性の判断を評価するものとは異なり，比較的微妙な関連性の順序を正確に判断できるか否かを表す指数であり，wgは全サンプルにおいてRel_A<Rel_Bとなった数（判断を誤った数）である．F_2はF_1と同様，値が大きくなるほど良い結果であることを表している．さらに，F_dはF_1とF_2を総合的に評価するための指数である．

このような，500組の尺度からランダムに100組を選択し，F_1, F_2, F_dを求める実験を1試行とし，全100試行について繰り返し行った．そして100試行により，各Nについて得られた100個のF_1, F_2, F_dを平均したものを実験結果とした．

4.4 実験結果と考察

実験結果を図4，図5，図6に示す．図4よりF_1についてはN=30で最大となり，両端付近では小さい値となった．これは取得する属性数が少ない場合では一致する属性も少ないため，十分なRel_Aを算出することができなかったことを表し，逆に取得する属性数が多すぎるとそれに伴って不適切な属性，つまり雑音属性が増加し，それら雑音属性同士が一致することによりRel_Cが大きくなり，F_1へ悪影響を及ぼしていると思われる．F_2については図5よりN=30，40でほぼ等しい最大値となったが，N=10の場合を除いて値に大きな変動はなかった．また，図6より，F_dについてはN=30で最大となったが，F_2と同様，N=10を除いて値に大きな差は見られなかった．しかし，F_1と同様にF_2, F_dにおいてもNの増加に伴って，取得される属性中に含まれる雑音属性の割合は増加する傾向が予想される．なお，N=0でF_d=0，N=∞でF_d=0となることは論理的に明確である．

よって，これらの考察からN=30が概念ベースにおけるより適切な固定属性数Nであることがわかった．

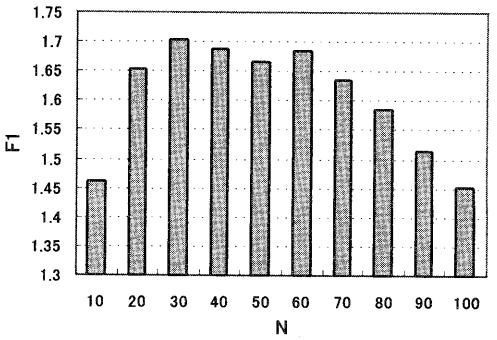


図 4:固定属性数評価実験結果（F_1）

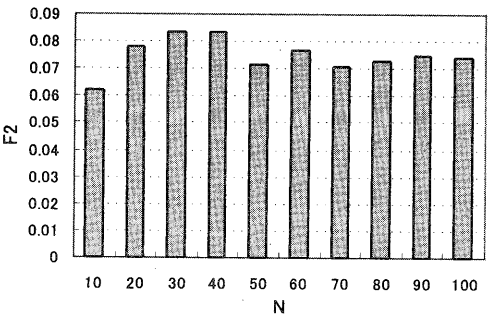


図 5:固定属性数評価実験結果（F_2）

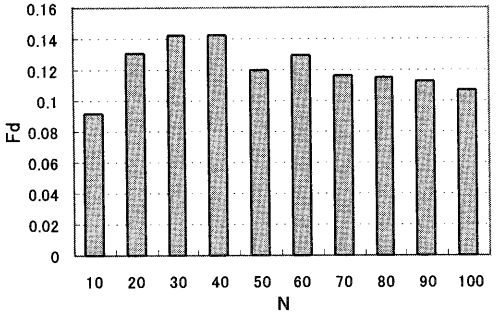


図 6:固定属性数評価実験結果(F_d)

5. 概念連鎖の関連度への影響

これまで用いた関連度計算方式は概念連鎖により，概念を2次属性まで展開し，2次属性の一致個数が最大となるように2次属性列を並び替え，2次属性の一致個数を考慮することによって関連度を算出するものであった．

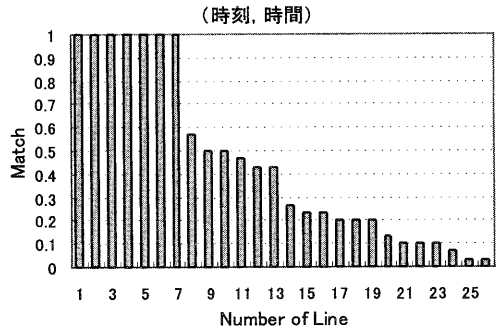


図 7:Rel(時刻, 時間)による一致度

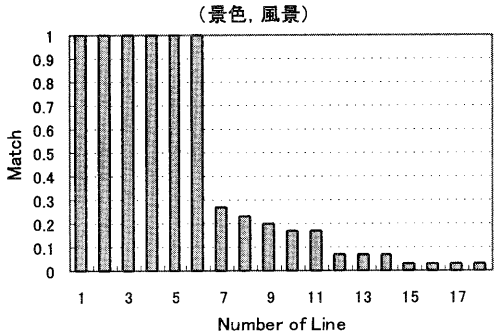


図 8:Rel(景色, 風景)による一致度

図7は概念“時刻”と“時間”の関連度を算出する過程での対応した2次属性列における一致度(対応する2次属性列における2次属性一致個数を0～1に正規化した数値)であり，また図8は概念“景色”と概念“風景”における一致度である(概念“時間”・“風景”は4.2において，基本概念“時刻”・“景色”に対し，それぞれRank_Aと定義した概念である)．

ここで，一致度が“1”となっている2次属性列は，全ての2次属性が一致した2次属性列であり，これらは1次属性の段階で完全一致したものであると解釈できる．よって，それら以外が1次属性の段階では一致を考慮されなかったが，新たに2次属性を取得することによって，ある程度の一致度を考慮されるべき，2次属性列といえる．前者をMT_F，後者をMT_Sとすると，図7，図8ともにMT_Sは最大でも0.6以下であり，MT_Fと比較して小さな値となっている．つまり，一致する2次属性数を数えることによって評価される一致度は関連度に対して十分な効果を与えていない可能性があるため，その影響力を考察する必要がある．

概念連鎖の関連度への影響を考察するために，まず，1次属性のみを用いて算出した1次関連度(case1)と，1次属性に加え2次属性まで展開して算出した従来方式による関連度(case2)を4.2で作成した評価尺度を用い，4.3と同様に500組の評価尺度からランダムに100組選択することによってF_1, F_2, F_dを求め，この試行を100回繰り返した100個のF_1, F_2, F_dの平均値を実験結果として，概念連鎖による2次属性の取得の関連度への影響を実験により評価した．

図9は実験結果であるが，F_1, F_2, F_dの全てにおいてcase2の方がcase1よりも優れた結果を示した．よって，概念連鎖による属性展開は有効であることが分かる．

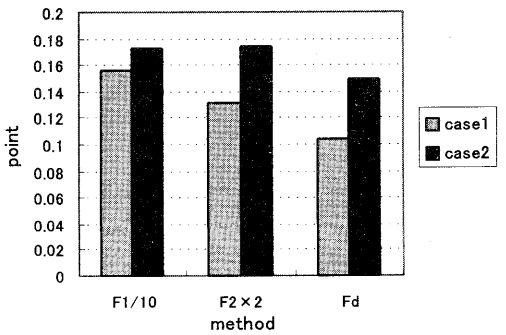


図 9:概念連鎖の関連度への影響

6. 変調関連度計算方式の提案

5 より，概念連鎖により，2 次属性を用いて関連度を算出することによって，より適切な結果を得られることが判明した．しかし，従来の関連度計算方式では，図 7，図 8 が示すように 2 次属性が関連度に対して強い影響を与えるものではないことは否めない．よって，その影響力をより際立たせることが可能となれば，さらに適切な関連度を与える計算方式となりうる．そこで，従来方式においては 2 次属性の一致個数は一致度に対して線形に反映するものであったが，新たに変調を行うことによって非線形に反映する方式（以降，変調関連度計算方式と呼ぶ）を提案する．

6.1 変調関連度計算方式 I（p 乗変調方式）

変調関連度計算方式として，2 次属性による属性一致個数を p 乗（p=1, 2, 3, 4, 5）し，関連度に反映させる方式を検討した．これは，全ての 2 次属性一致個数を p 乗することによって 2 次属性一致個数の関連度への影響を強めるものである．ここで，p の最大値が“5”であるのは，関連度計算に用いる固定属性数 N は N=30 であり，p 乗された数値が“30”を超える場合は一致度を“1”とするため，p が“5”以上の場合は全て同じ結果に収束するためである（p 乗の影響を受ける 2 次属性一致個数の最小は“2”であり，“2”の 5 乗は“32”となるため）．

この p 乗変調関連度計算方式を 4.2 で作成した評価尺度を用い，4.3 と同様に 500 組の評価尺度からランダムに 100 組選択することによって F_1, F_2, F_d を求め，この試行を 100 回繰り返した 100 個の F_1, F_2, F_d の平均値を実験結果として図 10 に示す．

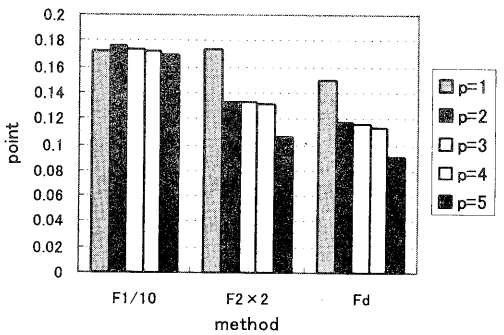


図 10：p 乗変調方式

図 10 より，F_1 についてはいずれの p においても，ほぼ同じ結果が得られたが，F_2 については p の増加に連れ減少した．これは p の増加により，Rel_A と Rel_B の正確な大小関係の判断が不可能になることを表している．よって F_1, F_2 を総合的に評価する F_d については p の増加につれて減少した．よって，p 乗変調方式は適切な変調関連度計算方式とは認められなかった．

6.2 変調関連度計算方式 II（シグモイド変調方式）

変調関連度計算方式を用いることにより，従来方式に比べ，より人間の判断に則した関連度を期待できるが，6.1 より，p 乗変調方式は p の増加により，F_2 が極端に減少するため，適切な変調関連度計算方式とは認められなかった．そこで，F_2 の増加，つまり比較的微妙な判断をより正確に行える変調関連度計算方式を見出すために，Rank_A, Rank_B それぞれに属する概念の基本概念との一致度の分布を調べた．

図 11 における $R(M)_{AB}$ は 500 個の Rank_A において，ある一致度が出現する総数 $(S(M)_A)$ を 500 個の Rank_B において，ある一致度が出現する総数 $(S(M)_B)$ で割ったもの(4)式であり，$R(M)_{AB}=1$（太実線）のとき，出現する一致度の総数が等しいことになる．

$$R(M)_{AB} = \frac{S(M)_A}{S(M)_B} \quad \cdots (4)$$

$$0 \le M \le 1$$

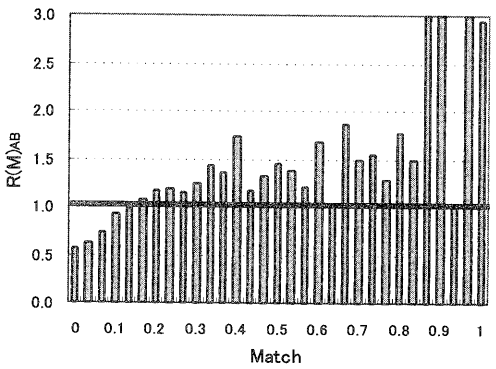


図 11：出現一致度比率

図 11 より，ばらつきはあるものの，一致度の増加に連れ，おおよそ大きな $R(M)_{AB}$ となっている様子がわかる．これは，基本概念に対し，Rank_A に属する概念の 2 次属性列は Rank_B に属する概念の 2 次属性列に比べ，

大きな一致度を持つ傾向があり，逆に Rank_B の 2 次属性列は Rank_A の 2 次属性列に比べ，小さな一致度を持つ傾向があることを表している．よって，小さな一致度ではより小さく，大きな一致度ではより大きく変化させることが可能となるような変調を行えば，より適切に Rank_A と Rank_B の大小関係の判断が可能になると思われる．そこで，このような変調を実現するために，(5)式のシグモイド関数(図 12)を基本形として 2 次属性一致個数の変調を行うこととした．

$$f(x)=\frac{1}{1+\exp(-x)} \quad \cdots \quad (5)$$

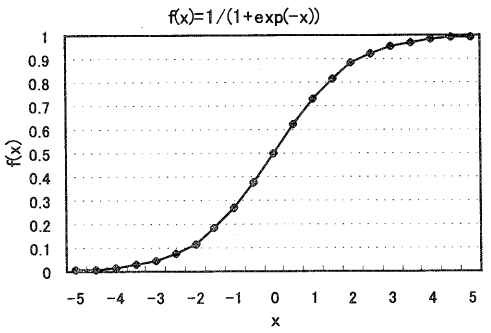


図 12:シグモイド関数

2 次属性一致個数を変調する基本関数として，(6)式を与える．(6)式において，右辺の分母を“30”としたのは，2 次属性一致個数の最大値を“30”とするためで，これにより，0~30 に収束するシグモイド関数となる．また，C(k)は第 k 列における 2 次属性一致個数であり，a, b は関数を決定する変数である．この変数 a, b を評価実験によって決定する．

$$f\{C(k)\}=\frac{30}{1+\exp[-a\{C(k)+b\}]} \quad \cdots \quad (6)$$

ここで，変数 a, b を決定する際，a, b が取り得る，ある程度の範囲を設定する必要があるが，今回は(7)式のような範囲とした．

$$\begin{aligned} 0.1 \le a \le 1 (0.01 step) \\ -20 \le b \le 20 (0.1 step) \end{aligned} \quad \cdots \quad (7)$$

変数 a, b を(7)式のような範囲とした理由は次に挙げる 4 つの理由によるものであり，いずれの場合においても得られる F_1, F_2, F_d は従来方式と比較して悪化した．

①a<0.1…得られる関数がほぼ線形関数となる
②a>1…C(k)の中間付近で急速に変調値が増大する
③b<-20…全ての変調値がほぼ最小値になる
④b>20…全ての変調値ほぼ最大値になる

よって，(7)の範囲において，4.2 で作成した評価尺度を用い，4.3 と同様に 500 組の評価尺度からランダムに 100 組選択することによって F_1, F_2, F_d を求め，この試行を 100 回繰り返した 100 個の F_1, F_2, F_d の平均値を実験結果とし，従来方式を用いることによって算出される F_d と比較して F_d の増加率が最も大きくなるような変数 a, b を決定した．

実験の結果，

$$a = 0.48$$
$$b = -11.6$$

のとき，従来方式に対する，シグモイド変調方式の F_d の増加率が最大となった．これらの変調定数を用いて決定される変調関数は図 13 のようになり，この変調関数を利用したシグモイド変調方式(sigmoid)を 1 次属性のみを用いた 1 次関連度計算方式(case1)，従来方式(case2)と比較すると図 14 のようになる．

図 14 より，F_1 についてはシグモイド変調方式を用いることにより，従来方式よりもやや小さい値となったが，これは大きな一致度がより大きく，小さな一致度がより小さくなるような変調を行ったため関連度のばらつきが大きくなり，標準偏差が増加したため，F_1 が小さくなったと思われる．しかし，F_2 については，従来方式に比べ，明らかに優れた結果が得られた．この結果から，シグモイド変調方式は従来方式に比べ，比較的微妙な判断をより正確に行える方式であることがわかる．また，F_d については F_1 の減少よりも F_2 の増加が大きく影響したため，case1, case2 に比べ優れた結果が得られた．よって，シグモイド変調方式は概念の 2 次属性までを評価する概念間関連度計算方式として有効な変調方式であることがわかった．

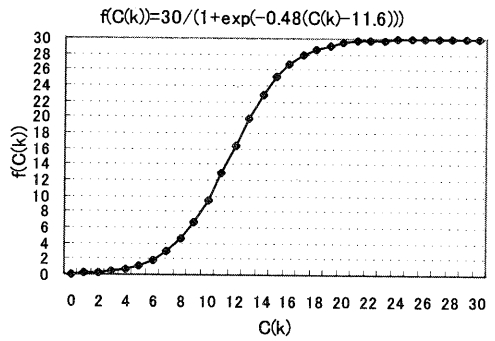


図 13:変調関数

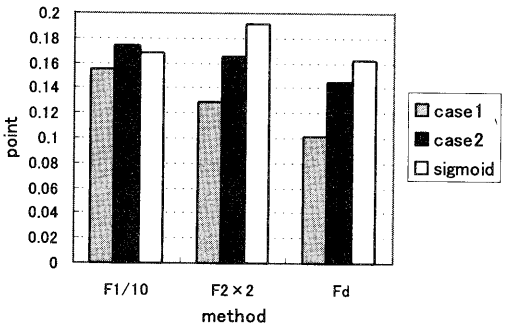


図 14:シグモイド変調方式

7. おわりに

知的判断メカニズムの実現には，その中核機構として，様々な判断を汎用的に行うための知識資源である概念ベースと概念間関連度計算方式を利用した連想機能を備える必要がある．ここで，概念ベースは維持管理の面から，可能な限り単純な構造を持つ必要があり，概念間関連度計算方式は人間の判断により則した，適切な計算方式であることが望まれる．

本稿では，概念ベースの各要素となる概念を自動学習や自動精錬の容易性の観点から，同一個数の属性で定義する場合，適切な固定属性数として N=30 が存在することを評価実験により示した．実際には具象名詞に比べ，抽象名詞や形容詞など，他の品詞における属性数は明らかに少ないが，これらについては，グループ分けを行い，各グループに対し，本稿で提案した方式を適用することが考えられ，今後実験による評価によってこれらを明らかにする．さらに，シグモイド関数を基本としたシグモイド変調方式による概念間関連度計算方式は，従来方式に比べ，より人間の判断に則した計算方式であることを評価実験により示した．

参考文献

[1] 入江 毅，渡部 広一，河岡 司，松澤 和光：知的判断メカニズムのための概念間の類似度評価モデル，信学技報，vol.98, No.499, AI98-75, pp.47-54 (1999)

[2] 笠原 要,松澤和光,石川 勉：国語辞書を利用した日常語の類似性判別,情報処理学会論文誌，Vol.38, No7, pp.1272-1283(1997)

[3] 石川 勉，井澤 潤次朗，Nguyen Viet Ha，笠原 要：単語の意味に関する概念ベースの類似性判別能力からの最適構成，人工知能学会誌，Vol.13, No.3, pp.470-479(1998)