E_033

概念属性の動的評価に基づく概念関連度計算方式

A Method of Degree of Association between Concepts based on Dynamic Evaluation of Attributes

奥村 紀之† 荒木 孝允† 渡部 広一† 河岡 司†
Noriyuki Okumura Takayoshi Araki Hirokazu Watabe Tsukasa Kawaoka

1．はじめに

人間は日常生活において，様々な連想をすることによって，幅広い会話やコミュニケーションを実現している．本稿では，人間の連想能力をコンピュータ上で実現するための連想メカニズムにおける動的な概念属性評価に基づく概念関連度計算方式，並びに，その応用として入力二語から常識的な連想を行うための二語連想メカニズムを提案する．

一般に連想機能として，ある語から別の関連の強い語を連想する機能と，同義語や類似語など意味の近い語を連想する機能の二つがある．前者の連想機能を想起語処理と呼び，後者の連想機能は未知の語を既知の語に置換する目的で使用されることが多いことから未知語処理と呼ぶ．すなわち，想起語処理は，「りんご」という入力に対して，「赤い，甘い，果物，…」のように人間が常識的に連想する関係の強い語を提示する機能である．一方，未知語処理は，会話など各種処理に必要な知識ベースに定義されていない語（未知語）が入力として与えられたとき，語の関連性を用いて，知識ベースに存在する語（既知語）に置換する処理である．

これらの処理を概念ベースと関連度計算方式によって実現する．関連度計算方式とは，語と語の関連の強さを概念の特徴の一致度合いから評価する手法である．たとえば，「赤ちゃん・子供」といった類似性の高い語のみならず，「赤ちゃん・おもちゃ」のように，類似ではないが関連の強い語同士の関係を定量化することが可能となる．

本稿では，概念属性を動的に評価することにより，より的確な関連性評価を行う関連度計算方式を提案すると共に，これを使った二語連想メカニズムを構築することによって，提案する関連度計算方式の有効性を示す．

2．概念ベース

概念ベース[1][2]は，常識判断メカニズム[3]，会話メカニズム[4]などを実現するための，連想メカニズムの中核となる大規模データベースである．概念ベースは，日本語の電子化辞書や電子化新聞を機械的に解析し見出し語を概念として説明文に含まれる自立語を属性として構成されている．概念ベースは連想メカニズムにおいて，関連度計算を行うための重要な要素となるデータベースである．概念ベースは概念(語)の集合であり，概念(A)は，それを特徴付ける属性(a_i)と，各属性(a_i)の重要性を示す重み(w_i)の対の集合(式1)によって構成される．

$$A=\{(a_1,w_1),(a_2,w_2),\cdots,(a_i,w_i),\cdots,(a_{znum},w_{znum})\} \quad (1)$$

概念ベースの各概念を特徴付けるための属性(a_i)もまた概念として定義されているため，一つの概念はn次元の属性連鎖集合によって定義される．

本稿では，電子化国語辞書から作成した3万語に電子化新聞から抽出した概念-属性を加え，拡張された約9万語の概念ベースを対象に評価している．電子化新聞から概念の属性を自動的に抽出する場合には，国語辞書から属性を抽出する場合に比べ，概念抽出の関係が希薄になり，より多くの雑音属性が含まれ品質は低下する．このため，より精度の高い関連度計算方式が必要となる．

3. 関連度計算方式

語と語の関連の強さを評価する手法として，関連度計算方式[5]が提案されている．関連度計算方式は，概念を重み付きの属性集合として定義し，それらの集合の一致度合いを関連度として計算する．属性集合の一致度合いを評価するため，まず，集合間で一致する，あるいは，類似する属性のペアを定義する．その後，定義されたペア属性に対し，各属性の重みを考慮し関連度を求める．

3.1 重み比率付き一致度

2つの概念A，Bに対し，その一次属性と重みをそれぞれ式2のように表現する．

$$A=\{(a_i,u_i)\mid i=1\sim L\}$$
$$B=\{(b_j,v_j)\mid j=1\sim M\} \quad (2)$$

このように概念が定義されているとき，概念A，Bの重み比率付き一致度$MatchWR(A,B)$を以下のように定義する．また，各属性の重みは，総和が1.0になるように正規化されている．

$$MatchWR(A,B)=\sum_{a_i=b_j} Min(u_i,v_j)$$
$$Min(u_i,v_j)=\begin{cases} u_i(u_i\le v_j) \\ v_j(v_j<u_i) \end{cases} \quad (3)$$

式(3)のように重み比率付き一致度を定義するのは，共通の属性を検出した際に，両方の属性の重みに共通する部分のみが有効に一致していると考えるためである．また，重み比率付き一致度は0.0～1.0の値を取る．

3.2 静的関連度計算方式

式(2)において，2つの概念A，Bのうち，属性の少ない概念を$A(L\le M)$とし，概念Aの一次属性の並びを重み順に固定する．

$$A=\{(a_1,u_1),(a_2,u_2),\cdots,(a_L,u_L)\} \quad (4)$$

このとき，まず，$a_i=b_j$ ($MatchWR(a_i,b_j)=1.0$)となる

† 同志社大学大学院工学研究科
Graduate School of Engineering, Doshisha University

属性を検索する．$a_i = b_j$ (完全一致)となる属性を検出した場合，$u_i > v_j$ ならば $u_i' = u_i - v_j$，$u_i < v_j$ ならば $v_i' = v_i - u_j$ とし，u_i'，あるいは v_i' を用いて他の属性と対応をとる．これにより，完全一致において対応に用いなかった重みを有効に利用でき，他の一致度の高い属性と対応させることが可能となる．そこで，式(2)の概念 A に対して，完全一致を考慮した後で，概念 B の並びを以下のように決定する．完全一致した属性が α 個あったとする．また，概念 B に関して，$B_{L+\alpha+1}, B_{L+\alpha+2}, \cdots, B_M$ は利用しない．

$$B_x = \{(b_{x_1}, v_{x_1}), (b_{x_2}, v_{x_2}), \cdots, (b_{x_{L+\alpha}}, v_{x_{L+\alpha}})\}$$
$$x_k = \{a_i と対応が決定した属性番号\} \quad (5)$$

このとき，$L+\alpha$ 個の属性が対応していることになるため，重み比率付き関連度 $ChainWR\ (A,B)$ を以下のように定義する．

$$ChainWR(A,B) = \sum_{i=1}^{L+\alpha} MatchWR(a_i, b_{x_i}) \times \frac{(u_i + v_{x_i})}{2} \times \frac{Min(u_i, v_{x_i})}{Max(u_i, v_{x_i})} \quad (6)$$

$$Min(u_i, v_{x_i}) = \begin{cases} u_i (u_i \le v_{x_i}) \\ v_{x_i} (v_{x_i} < u_i) \end{cases} \qquad Max(u_i, v_{x_i}) = \begin{cases} u_i (u_i \ge v_{x_i}) \\ v_{x_i} (v_{x_i} > u_i) \end{cases}$$

$ChainWR(A,B)$ は，$MatchWR(A,B)$ に各属性の重みの平均値 $(u_i + v_{x_i})/2$，並びに対応の決まった属性の重みの比率 $Min(u_i, v_{x_i}) / Max(u_i, v_{x_i})$ を掛け合わせることにより，関連度の補正を行う．$ChainWR$(机,椅子) を例に用いて重み比率付き関連度を解説する．机と椅子の一次属性を表 1 に，二次属性を表 2 に示す．

表 1.机と椅子の一次属性

概念	一次属性		
机	(学校,0.6)	(勉強,0.3)	(本棚,0.1)
椅子	(勉強,0.5)	(教室,0.3)	(木,0.2)

表 2.机と椅子の二次属性

一次属性	二次属性		
学校	(大学,0.4)	(校舎,0.4)	(木造,0.2)
勉強	(予習,0.5)	(試験,0.3)	(本,0.2)
本棚	(図書,0.6)	(書物,0.3)	(本,0.1)
教室	(教師,0.4)	(校舎,0.4)	(生徒,0.2)
木	(森林,0.5)	(木造,0.4)	(葉,0.1)

表 1 の机の一次属性・本棚と椅子の一次属性・勉強の重み比率付き一致度 $MatchWR$(本棚,勉強) は，式(3)，および，表 2 より，

$$MatchWR(本棚, 勉強) = Min(0.2_{本}, 0.1_{本}) = 0.1$$

となる．重み比率付き一致度は，完全に一致する属性の小さいほうの重みを加算することにより算出されるため，$MatchWR$(本棚,勉強) では，(本)のみが一致する属性となり，重み比率付き一致度は 0.1 となる．このことに基づき，すべての一致度の組み合わせを表 3 に示す．

表 3.一致度マトリックス

	学校	勉強	本棚
勉強	0	1	**0.1**
教室	**0.4**	0	0
木	0.2	0	0

ここで，(勉強)という属性は，椅子，机，双方に含まれる完全一致の属性となるため，$MatchWR$(勉強,勉強) = 1.0 となり，重みの差分(完全一致差分)を再び重み比率付き一致度の計算に利用している．式(6)，および，表 3 より，$ChainWR$(机,椅子) は，

$$ChainWR(机,椅子) = 1.0 \times \frac{0.3+0.3}{2} \times \frac{0.3}{0.3} + 0.4 \times \frac{0.3+0.6}{2} \times \frac{0.3}{0.6} + 0.1 \times \frac{0.2+0.1}{2} \times \frac{0.1}{0.2}$$

完全一致（勉強，勉強）　通常一致（学校，教室）　完全一致差分（勉強，本棚）

$$= 0.465$$

となる．静的関連度は，重み比率付き一致度と同様 0.0～1.0 の値を取る

3.3 動的関連度計算方式

静的関連度計算方式では，対象とする概念の属性のうち重みの高いものから s 個を抽出し，関連度計算に使用している．これは，概念の中には，最大で約 400 個(4.3 節)の属性を持つものもあり，各概念を特徴づける重みが大きいほどその概念を適切に特徴づける属性であるという観点から，(4.3 節)の結果から，重み順に上位 s 個の属性を用いて関連度を算出している．

一方，動的関連度計算方式では，重み順に属性抽出を行うのではなく，対象概念 A，B に共通する属性を優先的に抽出し，残りを重み順に抽出する．これにより，概念 A，B の共通属性が強調されるため，より人間の感覚に近い関連度が得られる．

例えば，式のように概念 A，B が定義されており，重み降順に属性が定義されているとする．この場合，重み順に上位三個の属性を使用し，関連度計算を行うと，共通する属性が全く存在しないため，関連度の値は 0 となる．しかし，本来は共通する属性が多数存在するため，静的に属性の対応を決める関連度計算方式では正しく関連度を算出できない．

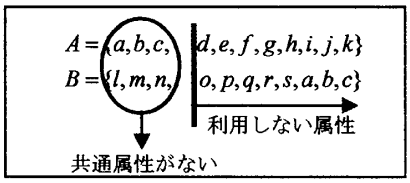

図 1. 正しく関連度が算出されない例

そこで，図 1 に示したに示すように，重み順に上位三個の属性を使用し関連度計算を行うのではなく，一致する属性を優先的に取得し，完全に一致するものを優先的に対応させた 3 個の属性を利用することにより，関連度を算出する(図 2)．すなわち，重み順に上位 3 個以下の属性との対応を考慮することにより，一致する属性の欠落を防ぐことが可能となる．

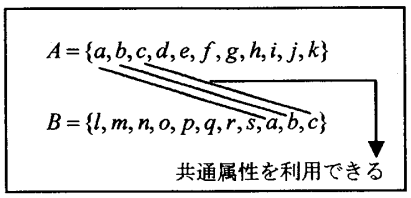

図2. 共通属性を優先的に利用する例

従って，この関連度計算方式では，計算対象となる概念ごとに使用する属性が動的に変化することとなる．

4. 評価

4.1 評価用データ

関連度計算の評価尺度として，4組1セットとなる $X-ABC$ 評価用データ(表 4)を用意する．任意の基準概念 X に対し，同義・類義等，最も関連が深いと考えられる概念 A，概念 X に対し，ある程度関連性が認められる概念 B，そして，概念 X に対してまったく無関連である概念 C によって構成する．このような評価用データを 1780 組用意した．

表 4.評価用データ

X	A	B	C
音楽	楽曲	音	電車
海	海洋	塩	車
学生	生徒	学業	林檎
…	…	…	…

4.2 評価方法

表 4 に示した，基準となる概念 X と対象となる概念 A,B,C の関連度をそれぞれ $ChainWR(X,A)$，$ChainWR(X,B)$，$ChainWR(X,C)$ とする．このとき，

$$ChainWR(X,A)-ChainWR(X,B)>AveChainWR(X,C)$$
$$ChainWR(X,B)-ChainWR(X,C)>AveChainWR(X,C)$$
$$AveChainWR(X,C)=\frac{\sum_{i=1}^{1780}ChainWR(X_i,C_i)}{1780}$$

を満たす時正解とする．関連度計算では，一つでも一致する属性が存在した場合，基準概念 X に対して，まったく無関連の概念 C であっても，関連度が 0.0 とはならず，ある程度誤差がある．そのため，$ChainWR(X,A)$，$ChainWR(X,B)$，$ChainWR(X,C)$ の間に，誤差以上の有意差が認められる場合のみを正解とする．この評価手法によって求められた割合を C 平均順序正解率と呼ぶ．

4.3 関連度計算方式の評価

静的関連度計算方式と動的関連度計算方式について，C 平均順序正解率を用いて評価した．各概念の属性数は，多いケースで約 400 個，最も少ないケースで 1 個である．そこで，重み順に上位 50 個を上限として静的関連度計算を評価した．また，動的関連度計算についても，同様に一致属性を最大で 50 個までとし評価した．図 2，図 3 に，静的関連度計算方式，並びに，動的関連度計算方式の C 平均順序正解率を重み順に上位 1 個から 50 個まで 1 個刻みで評価した結果を示す．

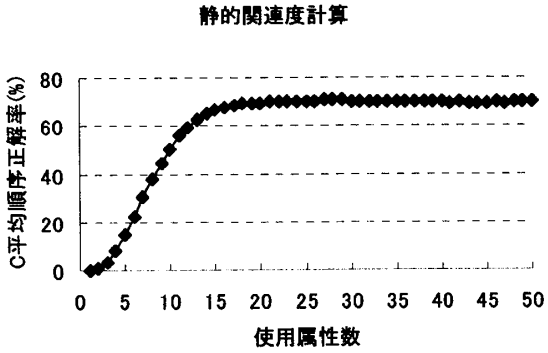

図 3. 静的関連度計算

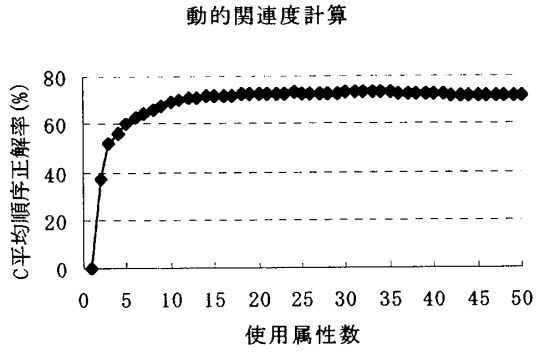

図 4. 動的関連度計算

図 3,4 に示したように，どちらの関連度計算方式を用いても，使用する属性数を増やすにつれ，C 平均順序正解率が収束傾向にあることが分かる．また，どちらの関連度計算方式においても，使用する属性数が 30 個のとき，C 平均順序正解率は最大となり，図 5 に示す値となる．さらに，概念ベースの各概念の属性数は平均約 30 個であることから，関連度計算に利用する属性数を 30 個とする．属性を 30 個使用した場合，C 平均順序正解率は図 5 に示すように，静的関連度計算方式では 70.39%，動的関連度計算方式では 73.2% となる．

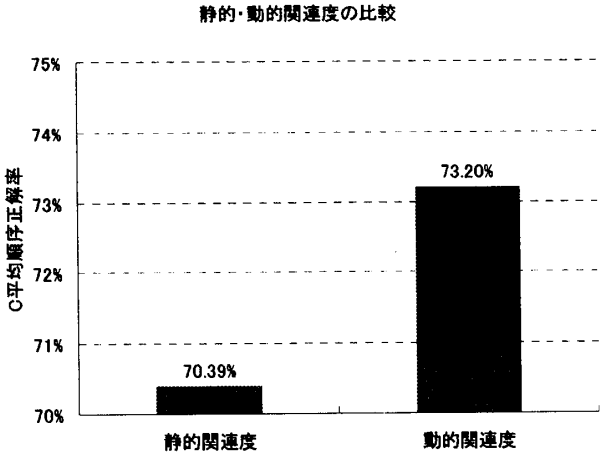

図 5. 静的・動的関連度の評価

また，表 5 に静的関連度，表 6 に動的関連度の *ChainWR*(*X*,*A*)，*ChainWR*(*X*,*B*)，*ChainWR*(*X*,*C*) の平均値を示す．

表 5.静的関連度計算方式の平均値

静的関連度		
ChainWR(*X*,*A*)	*ChainWR*(*X*,*B*)	*ChainWR*(*X*,*C*)
0.242 ⟷	0.048 ⟷	0.002
差 0.194	0.046	

表 6.動的関連度計算方式の平均値

動的関連度		
ChainWR(*X*,*A*)	*ChainWR*(*X*,*B*)	*ChainWR*(*X*,*C*)
0.302 ⟷	0.065 ⟷	0.002
差 0.237	0.063	

表 5 から，静的関連度よりも動的関連度のほうが *ChainWR*(*X*,*A*) の平均値と *ChainWR*(*X*,*B*) の平均値が高くなっていることが分かる．また，*ChainWR*(*X*,*C*) に関しては 0.002 のままであるため，一致する属性を優先的に利用する動的関連度を用いても，無関連概念間の関連度が正しく計算されていることが分かる．

次節に，連想機能を応用した二語連想メカニズムについて述べる．

5. 二語連想メカニズム

5.1 二語連想

本稿で提案している静的／動的関連度計算方式を用いて，(雨，長靴)など，二語から成る入力に対し，(傘，雨合羽，…)といった人間の感覚に近い連想を目的とした二語連想メカニズムを用いて提案方式の有効性を示す

二語連想メカニズムでは，概念ベースを参照し，入力された二語の 1 次属性，2 次属性，および，入力語を属性として持つ概念（1 次概念），さらにその概念を属性として持つ概念（2 次概念）を解答候補とし，連想語の絞り込みを行う．

多数取得された解答候補語と，入力二語との関連度の平均値を求め，関連度の平均値順に上位 10 個を連想語として出力とする．

5.2 二語連想の評価

二語連想メカニズムを評価するためテストセット(名詞，名詞)として，100 セット用意した．

評価方法として，3 人の人間が判定を行い，3 人すべてが正解だと回答したもののみを正解とした．各評価セットにつき，(正解語数)/(出力語数)を求め，100 セットの平均値を精度とする．また，アンケートによって各セットに対して，常識的であると考えられる連想語を収集し，各セットにつき，頻度順に上位 10 個を常識的な回答として，全評価セットで正解したもののうち，(常識的な回答に含まれる出力)/(100 セット×10 語)を再現率として評価した．

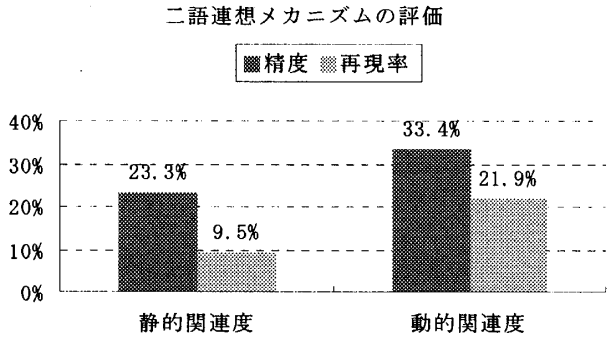

図 6.二語連想メカニズムの評価

図 6 に示したように，静的関連度計算方式から動的関連度計算方式に拡張することにより，精度，再現率共におよそ 10%の向上が見られた．しかし，二語連想メカニズムそのものの精度がまだまだ不十分であるため，概念ベースの精錬が必要であると考える．これは，概念ベースが辞書や新聞記事などから自動的に構築されており，適切でない属性が多数含まれているためである．

6. おわりに

本稿では，9 万語の概念ベースを使用する関連度計算において，概念属性を重み順に上位 30 個を利用して静的に使用する静的関連度計算方式と，計算対象となる概念が確定した段階で人間のように 2 つの概念に共通する属性を重視する，すなわち，使用する属性を動的に変更する動的関連度計算方式を提案した．また，これを単語から複数の語を想起する想起語処理に適用し，名詞二語の入力に対して想起語処理を行う二語連想において，動的関連度計算方式の有効性を実験により示した．

本研究は文部科学省から補助を受けた同志社大学の学術フロンティア研究プロジェクトの一環として行った．

参考文献

[1] 小島一秀，渡部広一，河岡司，常識判断のための概念ベース構成法-属性信頼度の考え方に基づく属性重みの決定，自然言語処理，Vol.9，No.5，pp.93-110，2002.

[2] 奥村紀之，渡部広一，河岡司，電子化新聞を用いた概念ベースの拡張と属性重み付与方式，情報処理学会研究報告，NL166，pp.55-62，2005

[3] 土屋誠司，奥村紀之，渡部広一，河岡司，連想メカニズムを用いた時間判断手法の提案，自然言語処理，Vol.12，No.4，pp.111-129，2005.

[4] 吉村枝里子，土屋誠司，渡部広一，河岡司 連想知識メカニズムを用いた挨拶文の自動拡張方式，自然言語処理，Vol.13，No.1，pp.117-141，2006

[5] 荒木孝允，渡部広一，河岡　司，共通・類似属性を考慮した概念間関連度計算方式　，情報処理学会第 68 回全国大会講演論文集，4N-2，2006