E-030

属性数に差のある概念間の関連度計算
A Method of Degree of Association for Concepts Differing in Number of Attributes

山村　伊織† Iori Yamamura
吉村枝里子‡ Eriko Yoshimura
土屋　誠司‡ Seiji Tsuchiya
渡部　広一† Hirokazu Watabe

1 はじめに

情報処理システムは，近年目覚しい発展を見せており，私たち人間社会のあらゆる分野で活用され，もはや欠かすことのできない存在となっている．しかし，こういった発展の多くは機能面，性能面におけるものであり，あらゆるユーザにとって親切であるとは言い切れない面もある．

利用者にとって使いやすい情報処理システムとは，高機能，高性能であることはもちろんであるが使い勝手の良さがきわめて重要である．そこで，人間どうしが会話をするのと同じように利用できる情報処理システムを開発することができれば，利用者はマニュアルを読んだり，操作手順を覚えたりする必要がなく非常に便利である．

人間が会話をする場合を考えると，あいまいな情報を受け取り適宜解釈することによって会話が成立していると考えられる．これは，人間が経験や学習によって獲得した言語やその基本となる単語概念に関する「常識」を持っているからである．すなわち，ある単語から概念を想起し，さらに，その概念に関係のあるさまざまな概念を連想できる能力が重要な役割を果たしていると考えられる．

そこで，コンピュータに常識的な判断を行わせるためには，コンピュータにも人間のように単語間の関連性を判断する連想メカニズムを持たせる必要がある．この連想メカニズムはある単語（概念）に対して関連の深い語（属性）並びに各属性の重要性を示す重みの集合によって定義された概念ベース[1]と概念と概念の関連の強さを定量化する関連度計算方式[2]によって実現されている．

関連度計算では概念の持つ属性を使用して関連度を求めている．既存の関連度計算方式では，関連度計算に使用する属性数を少ない方の属性数に合わせて関連度計算を行なっている．そのため属性数に大きな差のある概念間の関連度計算では属性数の多いほうの概念の属性はほとんど使用されない．これでは語の意味を十分に使用して関連度計算ができているとは言い難い．

そこで本稿では，輸送問題において計算される距離尺度である Earth Mover's Distance[3](以下 EMD)を適用し，概念の属性数に差が有っても語の意味を十分使用できる関連度計算方式を提案する．従来の関連度計算方式では，関連度計算に概念の意味をどれだけ使用するかは各概念の属性数で決まっていた．しかし，EMD を適用することで，関連度計算に使用する概念の属性数は，属性の重みの総和で決定することができる．そのため，EMD を適用した関連度計算方式を使用し，概念のもつ属性の重みを正規化することで属性の意味をすべて有効に使用することができると考えられる．

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

‡ 同志社大学理工学部
Department of science and engineering, Doshisha University

2. 概念ベース

概念ベースは言葉に関するデータベースで，ある概念 A は，その概念の意味を表す属性 a_i と，重要性をあらわす重み w_i の対で表される．

概念 A の属性数を N 個とすると，概念 A は次のように表せる．

$$A=\{(a_1,w_1),\ (a_2,w_2),\ \cdots,\ (a_N,w_N)\} \quad (1)$$

ここで，属性 a_i を概念 A の一次属性と呼ぶ．また，属性 a_i も概念ベースに登録されている 1 つの概念である．従って，a_i からも同様に属性を導くことができる．a_i の属性 a_{ij} を概念 A の二次属性と呼ぶ．概念「雪」を三次属性まで展開した様子を図 1 に示す．

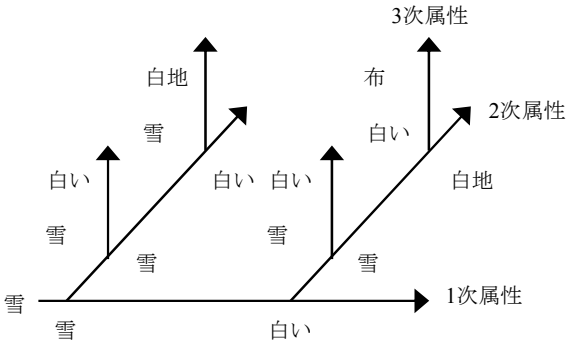

図 1：概念「雪」を概念ベースで属性展開した図

3. 関連度計算

概念間の関連度とは，2 つの概念 A と概念 B の関連の強さを定量的に評価するものであり，相対的な値である．例えば，概念「自動車」に対して，「車」，「自転車」，「馬」の関連の強さがどれほどのものかを知りたいとき，表 1 のように関連の強さを数値化できれば，コンピュータにも関連の強さの大小を判断できるようになる．この場合，概念「自動車」に対しては，「車」が最も関連が強いということがわかり，また「車」，「自転車」，「馬」の順に関連が強いこともわかる．

表 1：関連度の例

基準概念	対象概念	関連度
自動車	車	0.4
	自転車	0.18
	馬	0.02

概念間の関連の強さを定量的に評価する関連度計算方式には主に，概念間の共通属性を考慮して関連度を求める共通属性を考慮した関連度計算が用いられている[2]．本稿では，比較対象となる従来の関連度計算方式として共通属性を考慮した関連度計算を使用する．

3.1 一致度計算

一致度とは，概念の属性がどれだけ一致しているかにより関連の強さを定量的に評価する手法である．

概念 A，B をその一次属性を a_i，b_j，重みを u_i，v_j とし，属性がそれぞれ L 個，M 個($L \leqq M$)とすると

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

と表現でき，概念 A，B の一致度 $DoM(A,B)$は以下のようになる．

$$DoM(A,B) = \sum_{a_i = b_j} \min(u_i, v_j) \quad \min(\alpha,\beta) = \begin{cases} \alpha(\beta \geq \alpha) \\ \beta(\alpha > \beta) \end{cases} \quad (3)$$

（各概念の重みの総和は 1 に正規化する）

一致度は一致する属性のうち小さい方の重みの和となるが，これは両方の属性に共通して存在する重み分は有効であると考えるためである．なお，一致度は 0.0～1.0 の値を取る．

3.2 共通属性を考慮した関連度計算

共通属性を考慮した関連度 DoA(*Degree of Association*)は，対象となる二つの概念において，対象となる 1 次属性の全ての組み合わせについて一致度を求め，これを基に概念を構成する属性集合全体としての一致量を計算することで算出される．

具体的には，見出し語として一致する属性同士($a_i = b_j$)について，まず優先的に対応を決定する．他の属性については，全ての 1 次属性の組み合わせにおいて属性一致度を算出し，属性一致度の和が最大となるように組み合わせを決定する．一致度を考慮することにより，属性同士の見出し語としての一致だけではなく，一致度合いの近い属性を有効に対応づけることが可能となる．

また，概念 A，B 間の見出し語として一致する属性($a_i = b_j$)については，以下の処理により別扱いとする．$a_i = b_j$ なる属性があった場合，それらの属性の重みを参照し，$u_i > v_j$ となる場合は，a_i の重み u_i を $u_i - v_j$ とし，属性 b_j を概念 B から除外する．逆の場合は，同様に b_j の重み v_j を $v_j - u_i$ とし，属性 b_j を概念 B から除外する．見出し語として一致する属性が T 組あった場合，概念 A，B はそれぞれ A'，B'として以下のように定義し直され，これ等の属性間には見出し語として一致する属性は存在しなくなる．

$$A' = \{(a'_1, u'_1), (a'_2, u'_2), \cdots, (a'_{L-T}, u'_{L-T})\} \quad (4)$$
$$B' = \{(b'_1, v'_1), (b'_2, v'_2), \cdots, (b'_{M-T}, v'_{M-T})\} \quad (5)$$

見出し語として一致した属性の関連度を $DoA\_com(A,B)$ とし，以下の式で定義する．

$$DoA\_com(A,B) = \sum_{a_i = b_j} \min(u_i, v_j)$$
$$\min(\alpha,\beta) = \begin{cases} \alpha(\alpha \leq \beta) \\ \beta(\alpha > \beta) \end{cases} \quad (6)$$

次に，見出し語として一致する属性を除外した A'，B' の関連度を $DoA_def(A',B')$とする．$DoA_def(A,B)$を算出するために，属性数の少ない方の概念 A'の並びを固定し，属性間の属性一致度の和が最大になるように概念 B'の属性を並べ替える．この時，対応にあふれた属性は無視する．概念 A'の属性 a'_iと概念 B'の属性 b'_xが対応したとすると，概念 B'は以下のように並び換えられる．

$$B' = \{(b'_x, v'_x), (b'_{x+1}, v'_{x+1}), \cdots, (b'_{x+L-T}, v'_{x+L-T})\} \quad (7)$$

この結果，見出し語として一致する属性を除去した属性間の関連度 $DoA_def(A',B')$を以下の式によって定義する．

$$DoA\_def(A',B') = \sum_{s=1}^{x+L-T} DoM(a'_s, b'_s) \times \frac{\min(u'_s, v'_s)}{\max(u'_s, v'_s)} \times \frac{u'_s + v'_s}{2} \quad (8)$$
$$\min(\alpha,\beta) = \begin{cases} \alpha(\alpha \leq \beta) \\ \beta(\alpha > \beta) \end{cases}, \max(\alpha,\beta) = \begin{cases} \alpha(\alpha \geq \beta) \\ \beta(\alpha < \beta) \end{cases}$$

このように，見出し語として一致する属性間の関連度 $DoA\_com(A,B)$ と，それら以外の属性間の概念関連度 $DoA_def(A',B')$をそれぞれ算出し，合計を概念 A，B の関連度 $DoA(A,B)$とする．

$$DoA(A,B) = DoA\_com(A,B) + DoA\_def(A',B') \quad (9)$$

共通属性を考慮した関連度も，一致度と同様 0.0~1.0 の値をとる．また，実験より属性数は 30 個使用すればよいとの報告がなされている[2]ため，属性数は 30 個まで使用する．

3.3 関連度計算の問題点

既存の共通属性を考慮した関連度計算の課題として，使用する属性数として 2 つの概念で属性数の少ない方の数が関連度計算に優先されて使用されることが挙げられる．例えば，属性数 2 の概念 A と，属性数 30 の概念 B の間で共通属性を考慮した関連度計算を行なった場合に，使用しない属性を図 2 に示す．

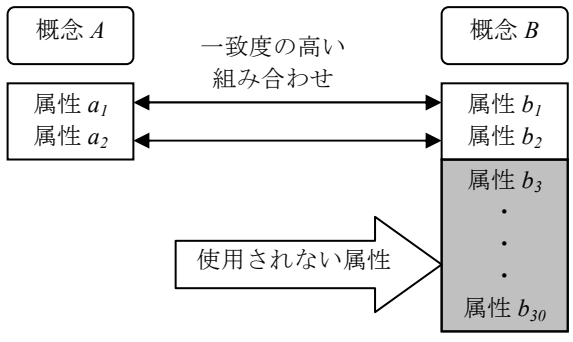

図 2：属性数に差のある関連度計算に使用されない属性

図 2 の場合では，概念 A の属性 a_1, a_2 とそれぞれ一致度の高かった概念 B の属性 b_1, b_2 のみが関連度計算に使用され，それ以外の b_3,から b_{30} までの概念 B の属性は関連度計算に使用されない．これでは語の意味を十分に使用して関連度計算ができているとは言えない．

そのため，本稿では概念の属性数に差が有っても語の意味を十分使用できる関連度計算方式として，EMD を適用した関連度計算方式を提案する．

4. EMD を適用した関連度計算

4.1 Earth Mover's Distance

EMD は線形計画問題の 1 つであるヒッチコック型輸送問題において計算される距離尺度であり，2 つの離散分布において，一方の分布を他方の分布に変換するための最小コストとして定義される．輸送問題とは，需要地の需要を満たすように供給地から需要地へ輸送を行なう場合の最小輸送コストを解く問題である．

EMD を求める際，2 つの分布は要素の重み付き集合として表現される．一方の分布 P を集合として表現すると，$P=\{(p_1,w_{p1}),...,(p_m,w_{pm})\}$ となる．今，分布 P は m 個の特徴量で表現されており，p_i は特徴量，w_{pi} はその特徴量に対する重みである．同様に，もう一方の分布 Q も集合として表すと，$Q=\{(q_1,w_{q1}),...,(q_n,w_{qn})\}$ となる．EMD の計算は，2 つの分布において特徴量の数が異なっている場合でも計算が可能であるという性質を持っている．ここで，p_i と q_j の距離を d_{ij} とし，全特徴間の距離を $D=[d_{ij}]$ とする．また，p_i から q_j への輸送量を f_{ij} とすると，全輸送量は $F=[f_{ij}]$ となる．コスト関数 $WORK(P,Q,F)$ は以下のように定義される．

$$WORK(P,Q,F)=\sum_{i=1}^{m}\sum_{j=1}^{n}d_{ij}f_{ij} \tag{10}$$

上記のコスト関数を最小とする輸送量 F を求め，EMD を計算する．ただし，上記のコスト関数を最小化する際，以下の制約条件を満たす必要がある．

$$f_{ij}\geq 0 \quad , \quad 1\leq i\leq m, \quad 1\leq j\leq n \tag{11}$$

$$\sum_{j=1}^{n}f_{ij}\leq w_{p_i}, \quad 1\leq i\leq m \tag{12}$$

$$\sum_{i=1}^{m}f_{ij}\leq w_{q_j}, \quad 1\leq j\leq n \tag{13}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}=\min\left[\sum_{i=1}^{m}w_{q_i},\sum_{j=1}^{n}w_{q_j}\right] \tag{14}$$

ここで，式(11) は輸送量が正であることを表し，p_i から q_j に送られる一方通行であることを表している．式(12) は輸送元である p_i の重み以上に輸送できないことを表す．式(13) は輸送先である q_j の重み以上に受け入れることができないことを表す．最後に式(14) は総輸送量の上限を表し，それは輸送先または輸送元の総和の小さい方に制限されることを表す．

以上の制約条件の下で求められた最適な全輸送量 F を用いて，分布 P，Q 間の EMD を以下のように求める．

$$EMD(P,Q)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}d_{ij}f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}} \tag{15}$$

4.2 関連度計算への適用

EMD は輸送問題の評価尺度であり，輸送問題は 1 つの供給地から複数の需要地へと輸送を行なうことが出来る．そのため，EMD を関連度計算に適用することで，1 つの属性が複数の属性と対応を取ることが可能になる．

EMD を関連度に適用するためには，EMD における分布を概念に，EMD における特徴量を概念の 1 次属性に，EMD における距離を 1 から重み比率付き一致度の値を引いた値に置き換える必要がある．さらに，このまま EMD を求めると概念 A , B が無関係の場合 1 に近い値が，関係性が強い場合 0 に近い値が算出されてしまうため，求められた EMD を 1 から引いたものを EMD を適用した関連度とする．

なお，概念の属性の重みの総和に差が有ると，関連度計算に概念の意味を十分に使用することはできない．そこで，EMD を適用した関連度計算では，概念の属性の重みの総和が 1 になるように重みの正規化を行なっている．

5. 評価実験

表 2 のような 4 つの概念を 1 組とする概念表記群を準備した．これを($X-A,B,C$)評価用データと呼び，任意の概念 X に対し，同義・類義など最も関連が深いと考えられる概念 A，概念 A ほどではないが関連があると思われる概念 B，そして完全に無関係である概念 C によって構成される．

このような性質を持つデータを人手によって 1780 組作成した．

表 2：($X-A,B,C$)評価用データの例

概念 X	概念 A	概念 B	概念 C
木刀	木剣	素振り	蔑称
時々	時たま	期間	議事

また属性数に差のある概念間の関連度計算において，EMD を適用した関連度計算方式の評価を行なうために，概念 X と概念 A，概念 X と概念 B の属性数に差のある概念の組を集めたテストセットを作成した．まず，($X-A,B,C$)評価用データの概念 X，A，B の属性数をそれぞれ $xnum$，$anum$，$bnum$ とし，概念 X と概念 A の属性数を比較し，属性数の多い概念の属性数を属性数の少ない概念の属性数で割ったものを $Zok(X,\ A)$とし，以下の式で定義する．

$$Zok(X,\ A)=\begin{cases} xnum/anum & (xnum>anum) \\ anum/xnum & (xnum<anum) \end{cases} \tag{16}$$

$Zok(X,\ A)$は概念 X と概念 A の属性数の比率を表している．この $Zok(X,\ A)$と $Zok(X,\ B)$が共に閾値 z を越える組を($X-A,B,C$)評価用データから集めた．閾値 z は 1.1 から 3.0 まで 0.1 刻みに設定することで，属性数に差のある概

念の組を集めたテストセットが 20 個完成する．以上の条件を式で表すと以下のようになる．

z=(1.1，1.2，1.3，…，2.9，3.0) (17)
$z < Zok(X, A)$ (18)
$z < Zok(X, B)$ (19)

以上の式を満たした$(X-A,B,C)$評価用データ 20 個と，$(X-A,B,C)$評価用データ 1780 組のテストセット 1 個の，合わせて 21 個のテストセットを用いて EMD を適用した関連度計算(以下提案方式)と共通属性を考慮した関連度計算(以下従来方式)の評価を行なう．

表 2 に示した概念 X と概念 A の関連度 $DoA(X, A)$，概念 X と概念 B の関連度 $DoA(X, B)$，概念 X と概念 C の関連度 $DoA(X, C)$が，

$DoA(X,A) > DoA(X,B)$ (20)
$DoA(X,B) > DoA(X,C)$ (21)

を満たすとき正解とし，テストセット中の正解の割合を順序正解率と呼ぶ．また順序正解率では，$DoA(X, A)$，$DoA(X, B)$，$DoA(X, C)$がほぼ同じ値でも，式(20)，(21)さえ満たしていれば正解と判断する．そこで，

$$Ave\ DoA(X,C)=\sum_{i=1}^{L} DoA(X,C)/m \quad (22)$$

$DoA(X, A) - DoA(X, B) > Ave\ DoA(X, C)$ (23)
$DoA(X, B) - DoA(X, C) > Ave\ DoA(X, C)$ (24)

を満たすときを正解とした割合を C 平均順序正解率と呼ぶ．ここで m は評価テストセット数である．

本来，概念 X と概念 C の関連度は 0 となるのが理想的である．しかし，関連度計算方式の特性上 2 つの概念間で 2 次属性までに共通属性が 1 つでも存在すれば，無関係な概念間でも一定の関連度が算出されてしまう．そこで C 平均順序正解率では，無関係な概念間の関連度の平均である $Ave\ DoA(X, C)$を誤差とみなし，誤差以上の有意な差が存在した場合を正解としている．関連度計算の評価尺度として C 平均順序正解率と順序正解率を使用する．

以上の条件で提案方式と従来方式の評価を行なった．C 平均順序正解率を図 3 に，順序正解率を図 4 に示す．

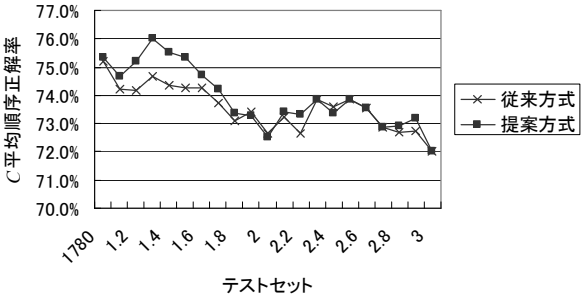

図 3：C 平均順序正解率

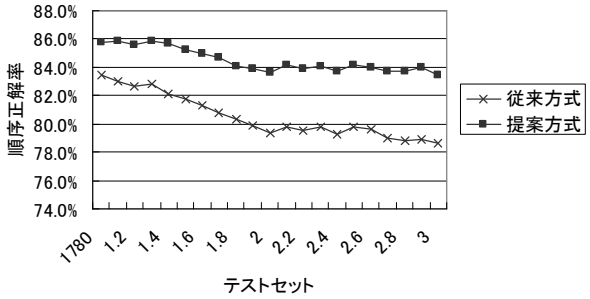

図 4：順序正解率

以上の結果より，提案方式の C 平均順序正解率は従来方式とほぼ同じであることが分かる．順序正解率は，概念間の属性数の差が大きくなるにつれて従来方式より提案方式が有効であることが分かる．C 平均順序正解率と順序正解率の違いは $Ave\ DoA(X, C)$を誤差として考慮するかどうかであり，$Ave\ DoA(X, C)$が大きくなると C 平均順序正解率は低くなる．以上より提案方式は従来方式と比べて雑音を多い関連度計算方式であると考えられる．

しかし，提案方式は従来方式と比べて C 平均順序正解率はほぼ同じ精度であり，順序正解率は提案方式の方が従来方式より精度が高い．以上より，雑音を多く拾ってしまうという点を考慮しても属性数に差のある概念間の関連度計算において，提案方式は有効であると判断できる．

6. おわりに

本稿では属性数に差のある概念間の関連度計算の精度向上を目的として EMD を適用した関連度計算を評価した．その結果，属性数に差のある概念間の関連度計算において有効であることが確認された．これは，1 つの属性が複数の属性と対応を取れるようになった結果，従来方式では使用できなかった属性を提案方式では使用できたため，正確な関連度計算が行なえたと考えられる．

7.謝辞

本研究の一部は，科学研究費補助金(若手研究(B)21700241)の補助を受けて行った．

参考文献

[1] 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，No. 7，pp. 1272–1283，1997.
[2] 荒木孝允，奥村紀之，渡部広一，河岡司，“比較対象概念の共通属性を重視する動的関連度計算方式”，同志社大学理工学研究報告，Vol.48，No.3，pp.14-24，2007.
[3] Y.Rubner，C.Tomasi，and L.Guibas，“The earth mover’s distance as a metric for image retrieval”，*Int.J.Comput.Vision*，Vol. 40，pp99–121，2000.