知的メカニズムのための概念間の類似度定量化方式

1M－6

入江 毅† 渡部 広一† 河岡 司† 松澤 和光††
†同志社大学工学部　††NTTコミニュケーション科学研究所

1. はじめに

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基本は知的メカニズムであり，その中核となる機構は概念ベースと概念の類似を利用した連想機能と考えられる．本稿では，概念ベース[2]を利用し，概念間の類似度を定量的に評価する方式を提案している．本方式の特徴は類似度を概念の2次属性集合で意味として評価している点である．また，4万語の概念ベースを使用した実験により本方式が人間の感覚により適合することを示した．

2. 概念の定義

ある概念Aは，その概念の意味特徴を表す単語の集合で表し，それらを概念Aの1次属性と呼ぶ．

概念 $A = \{ a_1^1, a_2^1, \ldots, a_i^1, \ldots, a_N^1 \}$

また，概念Aの1次属性はさらにその1次属性(概念）の意味特徴を表す単語の集合で表せる．これらの単語集合を概念Aの2次属性と呼ぶ．

1次属性 $a_i^1 = \{ a_1^2, a_2^2, \ldots, a_i^2, \ldots, a_M^2 \}$

このように，概念Aはn次属性まで定義可能であるが，本稿では2次属性までで概念が定義されているものとする．

3. 概念間類似度の評価モデル

2つの概念Aと概念Bの類似度を評価する場合，観点というものが必要になってくる．例えば，「飛行機」と「自動車」の類似度と「飛行機」と「すずめ」の類似度はどちらが大きいかを考えてみる．この場合，「乗り物」という観点で見る場合と「飛ぶ」という観点で見る場合とでは人間の感覚では類似度の逆転が生じるのが普通である．したがって，概念Aと概念Bの類似度は観点と呼ぶ第3の概念Cを指定したときに計算できるとするのが自然である．

しかし一般には，観点を陽に指定せずに2つの概念間の類似度を求めたい場合も多い．このような場合は，文脈やその他の状況から観点が陰に指定されていると考えられるが，そのような陰に指定されている観点を取り出すことは，現状では困難である．

Measuring Semantic Similarity Model between Concepts for Intelligent Mechanism
Takeshi Irie†, Hirokazu Watabe†, Tsukasa Kawaoka† and Kazumitsu Matsuzawa††
†Faculty of Engineering, Doshisha University
†† NTT Communication Science Laboratories

そこで，本稿では概念Aと概念Bの類似度を
(1) 観点が指定されていない場合
(2) 観点が指定されている場合
の2つの場合に分けて取り扱う．

3.1 観点なし類似度評価アルゴリズム

概念Aと概念Bの類似度を Sim(A,B) とし，アルゴリズムを以下のように定義する．

Ⅰ. 概念Aの1次属性の並びを固定する．

Ⅱ. 概念Bの各1次属性を対応する概念Aの各1次属性との一致度の合計が最大になるように並び替える（並び替わったBを B_x とする）．

$B_x = \{b_{x1}^1, b_{x2}^1, \ldots, b_{xi}^1, \ldots, b_{xN}^1\}$

ここで，一致度を次のように定義する．

◆一致度 $Match(a_i^m, b_j^m)$ の定義

2つのm次属性 a_i^m と b_j^m の一致度は，それぞれのm+1次属性の一致単語数を0から1の範囲に正規化したものとする．すなわち，

$a_i^m = \{a_1^{m+1}, a_2^{m+1}, \ldots, a_i^{m+1}, \ldots, a_L^{m+1}\}$
$b_j^m = \{b_1^{m+1}, b_2^{m+1}, \ldots, b_j^{m+1}, \ldots, b_M^{m+1}\}$

と表現し，$a_i^{m+1} = b_j^{m+1}$ なる a_i^{m+1} の個数をp個とするとき，一致度 $Match(a_i^m, b_j^m)$ を次式で定義する．

$$Match(a_i^m, b_j^m) = \frac{p(\frac{1}{L}+\frac{1}{M})}{2}$$

Ⅲ. 概念Aと概念 B_x の1次属性数を N_A, N_B とすると概念Aと概念Bの類似度 Sim(A,B) は先に定義した一致度を利用すると次式のように表せる．

$$Sim(A,B) = \frac{\frac{\sum_{i=1}^{N_A} Match(a_i^1, b_{xi}^1)}{N_A} + \frac{\sum_{i=1}^{N_B} Match(a_i^1, b_{xi}^1)}{N_B}}{2}$$

ここで，1次属性同士の並び替えは文献[1]で提案している「単純法」を用いている．この方法では一致属性数の合計が最大値になるとは限らないが，比較的最適解に近い値が出ることから，本稿で行う実験ではこの方法を採用した．なお，より最適な解を得たい場合は，遺伝的アルゴリズムなどを用いることができる[1]．

3.2 観点付き類似度評価アルゴリズム

概念Cを観点とする概念Aと概念Bの類似度を

Sim(A, B|C)と書くことにする．観点概念Cで概念Aと概念Bを比較するということは，観点概念Cに関係のある部分，すなわち観点概念Cが持つ1次属性に関係がある部分で評価するということである．アルゴリズムを以下のように定義する．

Ⅰ. 観点概念Cの1次属性の並びを固定する．

$$C = \{c_1^1, c_2^1, \ldots, c_i^1, \ldots, c_N^1\}$$

Ⅱ. 概念Aの各1次属性を対応する観点概念Cの各1次属性との一致度の合計が最大になるように並び替える．

$$A_x = \{a_{x1}^1, a_{x2}^1, \ldots, a_{xi}^1, \ldots, a_{xN}^1\}$$

Ⅲ. Ⅱと同様に，概念Bの各1次属性を対応する観点概念Cの各1次属性との一致度の合計が最大になるように並び替える．

$$B_y = \{b_{y1}^1, b_{y2}^1, \ldots, b_{yi}^1, \ldots, b_{yN}^1\}$$

Ⅳ. 概念Aと概念Bとの観点概念Cにおける類似度はA_xとB_yの（並びが固定された）各要素間の一致度と，観点概念CとA_xの各要素間の一致度，および観点概念CとB_yの各要素間の一致度の3者の相乗平均の（相加）平均とし，次式で表す．

$$Sim(A,B \mid C) = \frac{\sum_{i=1}^{N} \sqrt[3]{Match(c_i^1, a_{xi}^1) \times Match(c_i^1, b_{yi}^1) \times Match(a_{xi}^1, b_{yi}^1)}}{N}$$

4. 実験結果と考察

4.1 観点なし類似度の実験

図1は概念：警察とその概念の同義語ないしは類似語間での類似度計算の結果であり，濃い線が提案方式で薄い線がベクトルの内積による方式である．ベクトルの内積による方式とは，文献[2],「3」など従来の類似度計算方式であり，概念を1次属性の重みを要素とするベクトルとみなし，2つの概念の類似度を2つのベクトルのなす角の余弦で計算するものである．実験結果（図1）から（実際の実験はさらに多くのサンプルについて行っているが，ほぼ同様な結果を得ている），提案方式では従来の方式に比べ，10語の類似度の差が小さくなっている様子がわかる．

ここで，同義ないしは類似の関係にある概念を対象として類似度の計算を行ったわけであるから，人間の感覚と同じようにお互いの類似度の差ができるだけ小さくなることが望ましい．そういった意味から考察すると，提案方式によって得られた結果は評価できるものと考えられる．

4.2 観点付き類似度の実験

表1は3つの概念（看護婦，婦警，医者）について，2つの観点概念（女，病院）によって類似度を計算した結果である．表1より，2つの観点概念（女，病院）の両方について，人間の感覚と一致するような結果が得られたことがわかる．

このような実験を無作為に37通り行ったところ，2通りの観点概念が両方とも人間の感覚と一致しなかった例はなかった．また2通りともに人間の感覚と一致したのは19通りで，一方のみ一致したのは18通りであった．よって，本実験により提案した方式は十分評価できるものと考えられる．

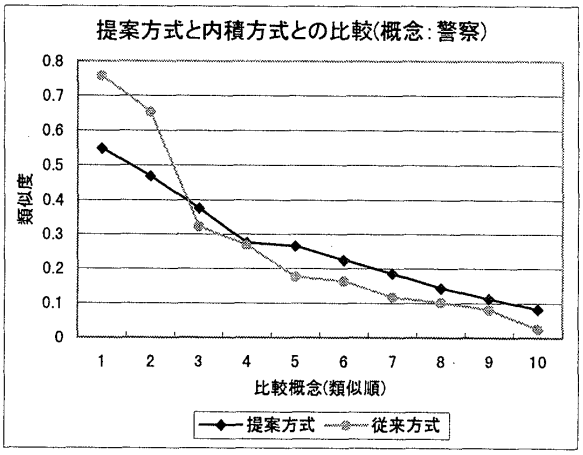

図1 観点なし類似度実験結果

表1 観点付き類似度実験結果

概念A	概念B	観点C	類似度
看護婦	婦警	女	0.142
看護婦	医者	女	0.038
婦警	医者	女	0.037

概念A	概念B	観点C	類似度
看護婦	婦警	病院	0.058
看護婦	医者	病院	0.222
婦警	医者	病院	0.054

5. おわりに

本稿では概念の意味を考慮し，概念の2次属性までを用いた概念間類似度の評価方式を提案した．今回の実験の範囲では提案方式により従来のベクトル内積方式に比べ人間の感覚により適合する結果が得られたが，今後それぞれの適用領域を明らかにする必要がある．

参考文献

[1] 浮田知彦, 渡部広一, 河岡 司：概念間の関連度計算への遺伝的アルゴリズムの適用, 情報処理学会春季全国大会, 1U-2(1998)

[2] 笠原 要, 松沢和光, 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(1994)

[3] Salton, G. and McGill, M. : Introduction to modern Information Retrieval, McGraw-Hill(1983)