E-39 概念ベースを用いた連想機能実現のための関連度計算方式

Measuring the Degree of Association between Concepts for Function of Assocoation with Concept-Base

井筒　大志†　渡部　広一†　河岡　司†
Daishi Izutsu　Hirokazu Watabe　Tsukasa Kawaoka

1.はじめに

人間の助けとなるような知的なコンピュータの実現には，人間の感覚と違和感の少ない常識的な判断を行わせる必要がある．常識的判断[1]とは，大きさ，速さ，場所，時間などといった量的なものに関する判断，赤い，丸いといった感覚に関する判断，うれしい，悲しいなどの感情に関する判断などのことをいう．こういった常識的な判断を行うことによって人間は，曖昧な情報を解釈し適切に会話を進めていくことができると考えられる．このような常識的な判断を実現させるためには，コンピュータに対して人間がある程度の知識を与える必要がある．しかし，全ての知識をあらかじめコンピュータに与えることは不可能である．そこで，限られた知識をより多くの語に対して適用するために連想を行う仕組みが必要である．連想とは，ある概念から関係のある他の概念を想起する機能のことである．

本稿では，この連想機能を実現するための 1 つの方法として概念ベース[2]を用いて語と語の関連性を数値化する方法[3]を提案する．語と語の関連性を調べる方法としては，ベクトル内積を用いる方法や属性集合の一致度で評価する方法，語の表記特徴を利用する方法が知られている．このうち，属性集合の一致度で評価する方法については，概念ベース内での重みの扱いと属性の一致のさせ方が問題となる．そこで，属性集合の一致度で評価する方法について属性の一致のさせ方と重みの扱い方，また概念ベースの評価方法について検討を行う．

2.概念ベース

概念ベースは，複数の辞書から機械的に構築された日常使う語に関する大規模で汎用的なデータベースである．しかし，概念ベースは機械的に構築されているため，必ずしも適切なデータのみで構成されているわけではない．そこで，このような不適切なデータを削除（概念ベースの精錬）する必要がある．本稿では，精練された概念ベースを利用して実験を行った．

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 は以下のように表せる．

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

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．概念ベース中では表 1 のように概念が格納されている．

本稿で実験に利用した概念ベースには，約 9 万の概念があり，属性数は概念ごとに異なる．1 概念あたりの平均属性数は約 29 個である．また，重みは情報量を利用して付与されている．

表 1　概念の例

概念	属性
電車	（客車,115）（貨車,105）（汽車,104）・・・
画像	（肖像,81）（映像,74）（画面,43）・・・
野球	（球技,85）（投手,76）（走塁,52）・・・

3.関連度計算方式

二つの概念 A，B 間の関連度は，それぞれの概念を二次属性まで展開し，一致する属性とその重みによって求める一致度によって計算する．一致度は 0～1 の実数値をとり，属性がすべて一致する場合が 1 である．具体的には，一致する二次属性を調べその重みを使って計算する一致度の和が最大になるような一次属性の組み合わせを作る．一次属性の組み合わせを作る場合には，GA などを使うことによって最適な組み合わせを作ることが考えられるがここでは，最大値を取る組み合わせを順に取っていくことによっても比較的良い関連度が得られるものと考える．

2 つの概念 A，B の関連度を rel(A,B)と定義し，関連度の計算方法を示す．

3.1 一致度の計算方法

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<M)とすると

$$A=\{(a_1,u_1),(a_2,u_2),\cdots,(a_L,u_L)\}$$
$$B=\{(b_1,v_1),(b_2,v_2),\cdots(b_M,v_M)\}$$

と表現する．概念 A，B の一致度 MatchW(A,B)は

$$MatchW(A,B)=\sum_{a_i=b_j}\min(u_i',v_j')$$

と定義する．このとき，一致度は一致する属性のうち小さい方の重みとなるが，これは両方の属性に共通して存在する重み分は有効だと考えるためである．ただし，重みは概念ごとに固有なので，各概念ごとに重みの総和が 1 になるように正規化しておく(u',v').

3.2 関連度の計算方法

関連度は，計算対象となる概念の全ての一次属性の組み合わせについて一致度を計算し，一次属性どうしの対応を決定することによって計算する．具体的には，一致する一次属性どうしについては優先的にその対応を決定する．一致しない部分については，その一致度の合計が最大になるように一次属性どうしの対応を決定する．一致度を利用することによって，完全に一致しない一次属性についても関連の度合いを考慮に入れることができる．

一致する一次属性の組み合わせがあれば，その一次属性の組み合わせについて，以下の式のように，一致度（属性

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

が完全に一致するので1）に対応する属性の重みのうち，共通する分を掛ける．

$$rel(A,B)=\sum_{a_i=b_j}1\times\min(u_i',v_j')$$

また，一致する一次属性どうしで対応が取れた場合には，元の重みからそれぞれ一致度に掛けた分の重みを引く．重みを引いたあと，重みが 0 でなければ，その属性はさらに他の属性と対応をとることができる．

$$u_i''=u_i'-\min(u_i',v_j')\qquad v_j''=v_j'-\min(u_i',v_j')$$

一次属性どうしが一致しない場合，一致度が高いものから順に一次属性の組み合わせを作るために概念 A の属性の並びを固定し，概念 B の属性を並び替える．

$$A=\{(a_1,u_1),(a_2,u_2),\cdots,(a_L,u_L)\}$$
$$B=\{(b_{x1},v_{x1}),(b_{x2},v_{x2}),\cdots(b_{xL},v_{xL})\}$$

属性数は少ないほうに合わせる．このように作った組み合わせについて次式のように関連度を計算する．

$$rel(A,B)=\sum MatchW(a_i,b_{xi})\times(u_i+v_{xi})\times(\min(u_i,v_{xi})/\max(u_i,v_{xi}))/2$$

全体を 2（正規化した重みの合計）で割るのは関連度が 0～1 の値をとるようにするためである．一致度と関連度の計算例を概念「机」と「椅子」を例に示す．2 つの概念の一次，二次属性を表 2，3 に示す．

表 2　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 3　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)

一次属性「勉強」と「本棚」の一致度は次式のように計算できる．

MatchW(勉強,本棚)=*min*(0.2,0.1)=0.1

同様に全ての一次属性の組み合わせについて一致度を計算した結果が表 4 である．

表 4　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

関連度の計算は，まず完全一致部分から行う．次に残りの部分から一致度の大きいところから順に対応を決めると関連度は次式のように計算できる．

$$rel(\text{机，椅子})=1\times0.3+0.4\times(0.3+0.6)\times(0.3/0.6)/2+0.1\times(0.2+0.1)\times(0.1/0.2)/2$$

4.評価実験

表 5 のような 4 つの概念の組を用意する．ここで，概念 X は任意の概念（対象概念）であり，概念 A は概念 X と同義または類義の概念，概念 B は概念 X に密に関連する概念，概念 C は概念 X に疎な概念である．すなわち，対象概念 X に対して A が非常に関連が強く，B は関連があり，C はほとんど関係がない概念である．

表 5　評価用データ

概念 X	概念 A	概念 B	概念 C
椅子	腰掛け	机	像
医師	医者	看護婦	山
飲料	飲み物	喉	反省

表 5 の各評価用データの組に関して

$$rel(X,A)-rel(X,B)\geq\varepsilon$$
$$rel(X,B)-rel(X,C)\geq\varepsilon$$

を満たせば，その関連度計算結果を正解とし，それ以外は不正解である．このような評価用データを人手で作成し，3 人中 3 人が正しいと判断した 590 組中何組のデータが正解したかによって関連度計算の評価実験を精錬を行った概念ベース[2]を対象に行った．評価実験の結果は図 1 のようになった．

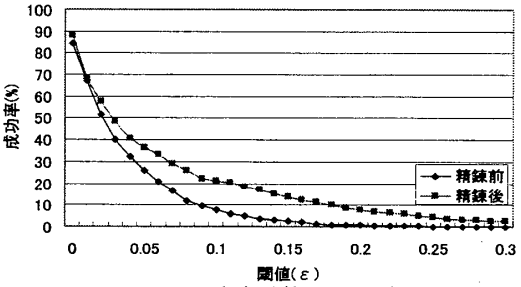

図 1　関連度計算の評価結果

ここで閾値 ε の値と評価用データの成功率が問題となる．ε =0 のとき，精錬前後の概念ベースで成功率には大きな差はない．しかし，0.05≦ ε ≦0.2 の範囲では，精錬後の概念ベースの方が明らかに成功率が高くなっている．すなわち，関連度と閾値を利用することにより，概念ベースを正しく評価できることがわかる．

5.おわりに

本稿では，常識的判断メカニズムの実現のために重要であると考えられる連想機能実現の方法の一つとして，概念間の関連性を数値化する方法を提案した．提案手法では，属性の質の異なる二つの概念ベースの評価を行うことが可能であり，概念間の関連性を正確に数値化する有効な手法であると考えられる．

謝辞

本研究は文部科学省から助成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「知能情報科学とその応用」における研究の一環として行った．

参考文献

[1] A.Horiguchi, S.Tsuchiya, K.Kojima, H.Watabe, T.Kawaoka (2002). Constructing a Sensuous Judgment System Based on Conceptual Processing, Computational Linguistics and Intelligent Text Processing (Proc. of CICLing-2002), Springer, pp.86-95

[2] 広瀬幹規,渡部広一,河岡司(2002).“概念間ルールと属性としての出現頻度を考慮した概念ベースの自動精錬手法” 信学技報,TL2001-49,pp.109 –116

[3] 渡部広一,河岡司(2001).常識的判断のための概念間の関連度評価モデル,自然言語処理,Vol.8,No.2,pp.39-541