

概念間の関連度計算への遺伝的アルゴリズムの適用

1 U - 2

浮田知彦 渡部広一 河岡司

同志社大学 工学部 知識工学科

1. はじめに

コンピュータに人間的な柔らかい判断をさせるためには、まず、意志表現の基本となる語概念の間に人間が持っているのと同じような語の関連を見出す仕掛けを持たせることが必要である。これについては、従来、シソーラス[1]を使った概念知識データベースを構築する方法が種々研究されている[2]。本研究では、今後出てくる新語も含め、40万語以上と言われている全ての語を対象に語の関連を定量的に表す方法について提案する。基本的にはコンピュータにも、人間の場合と同様、先生、友達、辞書、辞典、書物などが与えてくれる知識をそのまま利用して未知語の関連度を求める。具体的には全ての語を関連語の集合（マトリックス）として表現し、このように表現された2つの語の関連度を遺伝的アルゴリズム（Genetic Algorithms、以下GA）を適用することにより数値化する手法を提案する。

2. 語概念のモデル化と関連度

コンピュータに関連度を計算させるには、関連度計算したい語（以下対象語）をモデル化する必要がある。その一つの方法として、対象語と関係する語（以下要素語）を並べて横ベクトルとし、その並べた語同士が全体のどれほどの割合で一致するかを調べる方法が考えられる。しかし、この方法では片方の語に「丸い」、もう片方の語に「円形」と言う語が含まれていても全く関係のないものと判断してしまう。要素語に柔軟性を持たせるため、要素語を列ベクトルとして、すなわち、元の対象語を要素語2世代のマトリックスで表現する。このマトリックスを対象語の語概念としてモデル化する。要素語からさらに関係する語（以下属性語）を並べてマトリックスとする方法である。すなわち、このモデル化により、2つの対象語の関連度を語概念マトリックスの近さ(G)で定義する。2つの語のマトリックスサイズは必ずしも同じである必要はないが実験ではマトリックスのサイズを 20×20 とした。

3. マトリックス概念による関連度計算法

生成されたマトリックスから、関連度 (G) を求め

Measure of Word-Relation Calculation between Concepts by Genetic Algorithms

Tomohiko UKITA, Hirokazu WATABE, Tukasa KAWAOKA
Doshisha University

る方法として次の3つの方法がある。

- i) 要素語のみの一致数を調べる。
 - ii) マトリックス全体を一塊りとして、要素語、属性語の一致数を調べる。
 - iii) 要素語（列ベクトル）を最適な配置に並び替えて対応する列同士の一致度を調べる。
- i) は2項の概念のモデル化で述べたように情報量が少なすぎて柔軟な判断に欠ける。
 ii) は要素語の属性語が並べてある列ベクトルをばらばらにするため対象語の概念を変えてしまう。
 iii) の方法を、具体例を図1に示し詳しく説明する。図1のように「自動車」と「飛行機」の要素語が与えられたとする。ここで「自動車」の要素語の並びは固定しておき「飛行機」の要素語を「自動車」の要素語と出来るだけ対応がよいように並べ替えをするという方法である。

この要素語の並べ替えの際に属性語の一致数を使用する。すなわち、一致する属性数が多い要素語同士がよい対応であると見なす。

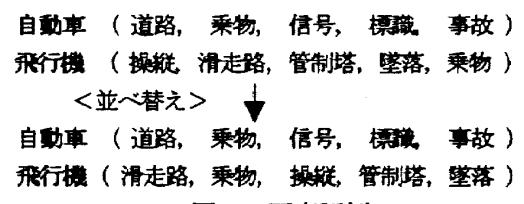


図1. 要素語例

対象語 A (自動車), B (飛行機) のマトリックスから各要素語間の関連度を表した X マトリックスを生成する。X マトリックスとは、A,B のそれぞれの各列ベクトルとの一致属性語数を並べたものである。A の第1要素語と B の1第一要素語との関連度を X マトリックスの (1,1) に、A の第一要素語と B の第二要素語との関連度を X マトリックスの (1,2) として生成したものである。X マトリックスから最高値を求めるには全組合せを計算すればよい。しかし、本実験マトリックスサイズの 20 を例にしてみると、組合せは $20!$ (2.4×10^{18} 通り以上) 存在する。そのすべてを計算する事は实际上不可能と言える。

しかしながら、元々の語概念の関連度と言う物理的な意味から考えれば、必ずしも最高値である必要はなく、準最高値で十分であり、これを求めるには次の3つの方法が考えられる。

Ⅲ-1) ランダムサーチ法

Ⅲ-2) 単純法

Ⅲ-3) GA

ランダムサーチ法は、文字通りランダムに X マトリックスから要素を選出し、より高い数値を選んでいく方法であるが、現実的な試行回数では精度が得られない。

単純法は、X マトリックスから最高値を選出し、次にその数値の同一行・列以外から最高値を選出していくという、最高値に注目した大まかで単純な方法である。この方法では、計算時間も短くかなりよい値が出るようと思われるが、図 2 に示す例（数値による）の様に明らかに次の GA より大きく劣る例が存在することがわかる。

A					B				
1	10	6	51	52	1	2	8	46	53
2	11	7	72	51	2	1	23	47	52
3	30	8	73	61	3	20	24	43	51
4	31	9	74	62	10	21	25	61	54
5	32	10	63	63	11	22	26	62	63

X	3	2	0	0	0
	2	0	0	0	0
	1	0	1	0	0
	0	0	0	0	2
	0	0	0	2	3

単純法 = 28%
GA = 37%

図 2. (数値による) マトリックス例

4. 遺伝的アルゴリズムの適用

本研究では、X マトリックスを遺伝子と見なし、各行・列から一つづつしか遺伝子を取らないという制約を加えて逆位・エリート保存を行った。

1) 逆位

図 3 のように遺伝子を一組又は複数入れ替えて、次世代の個体を生成する方法である。ただし、 x_i は X マトリックスの各要素である。

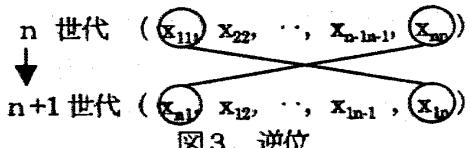


図 3. 逆位

2) エリート保存

次世代に進化したときに優秀な個体が退化してしまうのを防ぐために、優秀な個体上位数個を次世代にそのまま残す方法である。

5. 実験結果と考察

実際は概念マトリックスに単語が並べられているが、実験では、図 2 に示したように単語の代用として数値を利用することにより計算の精度を確認した。乱数の幅を変化させることにより、 x_i が一致する確率を変化させてデータを採取することにした。その結果、図 4 に示すように単純法と GA を比較した結果、同じ値が出ることはあっても、GA の方が劣るデータが 1 つもなかった。（実際には GA 適用の際の初期個体として単純法で求めた解を入れておくことにより保証できる。）また、「飛行機」と「自動車」の結果も図 4 中に示し、図 5 に実際のマトリックスの一部抜粋した例を示す。

上記、考察と結果から、GA を適用することにより現実的な時間内でより精度の高い関連度求めることが出来る。

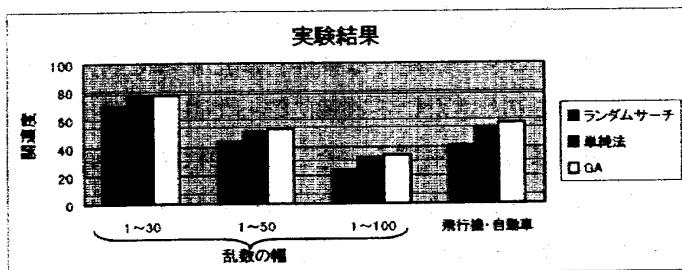


図 4. 実験結果

飛行機				自動車			
飛行機	操縦	空港	滑走路	自動車	道路	渋滞	信号
操縦	機長	旅行	離陸	道路	交通	事故	運転
:	:	:	:	:	:	:	:
滑走路	運転	税関	地上	渋滞	工事	帰省	無視
管制塔	信号	免税	誘導	信号	高速	旅行	警察

図 5. 飛行機・自動車のマトリックス

6. おわりに

本稿では語概念間の関連度計算の手法として GA を適用する方法の有効性を示した。今後は、実際的な類似語の自動抽出に置いて、計算時間の短い単純法で候補を絞り、最終的には GA 適用により精度を上げなどの適用法について検討する。

参考文献

- [1] NTT コミュニケーション科学研究所：日本語語彙大系、岩波書店（1997）
- [2] 笠岡、松澤、石川、河岡：観点に基づく概念間の類似性判別、情報処理学会論文誌、Vol.35, No.3, pp.505-509 (1994)