

知的判断メカニズムのための概念間の類似度評価モデル

入江 毅[†] 渡部 広一[†] 河岡 司[†] 松澤 和光^{††}

[†] 同志社大学工学部知識工学科

〒610-0394 京都府京田辺市多々羅都谷 1-3

^{††} NTT コミュニケーション科学研究所

〒619-0237 京都府相楽郡精華町光台 2-4

あらまし 柔軟な利用者インタフェースをも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。概念ベースは維持管理の面からできるだけ単純な構造でなければならない。本稿では、概念ベースに課せられたこのような条件を前提に概念間の類似度の定量化方式を提案している。本方式の特徴は類似度を概念の1次属性のベクトルの相違としてではなく概念の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。

キーワード 知的メカニズム, 概念ベース, 概念連鎖, 類似度, 観点, 1次属性, 2次属性

Measuring Semantic Similarity Model between Concepts for Intelligent Mechanism

Takeshi Irie[†], Hirokazu Watabe[†], Tsukasa Kawaoka[†] and Kazumitsu Matsuzawa^{††}

[†] Department of Knowledge Engineering and Computer Science, Faculty of Engineering,
Doshisha University, Tanabe, Kyoto 610-0394

^{††} NTT Communication Science Laboratories

2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan 619-0237

Abstract The basic factor of information processing systems which have friendly user interfaces is a intelligent mechanism, and the main elements of that are the concept-base and the association mechanism using similarity between concepts in the concept-base. A structure of the concept-base must be simple since the concept-base can be maintained easily. This paper proposes measuring semantic similarity model between concepts based on such a conditions of the concept-base. The feature of proposed model is not that the degree of similarity evaluates by difference of first stage attributes vector but the degree of similarity evaluates by set of second stage attributes as semantic points. By experimental results using the concept-base which consists of about 40,000 concepts, it is shown that the proposed model can output the closer degree of similarity to that decided by human judges.

key words concept-base, chain concept, measuring semantic similarity, view point, n-stage attribute

1. はじめに

情報処理システムは、ここ数年を振り返ってみても従来にもまして急速な発展を続けており、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になっている。しかし、これらの発展は、ほとんど、機能面、性能面での高度化に起因するものであり、知的な観点での著しい進展はみられない。今後、情報処理システムの機能・性能面での高度化が進めば進むほど、知的な情報処理へのあこがれは一層強まってくるものと思われる。「知的」や「知能」については古くから哲学、生物学、認知科学、理学、工学の幅広い領域で各種の研究が行われてきた。工学的な観点での「知的」に限っても、人工知能の実現可能性をめぐる多くの議論、研究がなされてきているが、未だに「知的」に関する明快な定義さえ与えられていない。

我々の研究のねらいは、このように難しく、やっかいな「知的」の本質の解明に挑戦しようとするものではなく、知的と呼んでもあまり違和感がなく、しかも、従来とは異なるメカニズムで、情報処理の高度化につながるような現実的なメカニズムを創出することにある。これまで、コンピュータ技術をベースに、より知的なメカニズムを求め、ゲーム理論や知識処理など人工知能と呼ばれる分野を中心に、自然言語処理、ファジィ、ニューラルネットワーク、遺伝的アルゴリズム、人工生命などの研究が行われ、それぞれに著実な成果を積み上げてきている。しかし、これらの多くは、基本的には、従来の情報処理技術とは比較的独立な技術として、あるいは、情報処理の特定の応用領域への適用を対象に研究が行われてきた。

本研究で対象とする「知的判断メカニズム」は、コンピュータシステムにおいて、オペレーティングシステム（OS）がハードウェアの効率的で使い易い利用メカニズムを汎用的に提供しているように、コンピュータシステムに「連想概念ベース」と呼ぶある種の知識をコンピュータ資源として持たせ、「知的メカニズム」として、これの利用メカニズムを汎用的に提供することをねらうものである。データベース、ワードプロセッサ、帳票処理、インターネットなどの情報処理システム（プログラム）の開発において、OS機能と同様、これらの知的判断メカニズムを適宜に使うことにより、柔らかなユーザインタフェースをもち、利用者の意図も汲み取ることのできる、「知的」と呼んでもよいような情報処理システムを開発することができると考えている。

本稿では、この知的メカニズムの基盤となる「連想概念ベース」を用いた概念間の類似度の定量化方式について提案している。概念類似度のテーマについては、幾つかの興味ある研究成果が報告されているが[1],[2],[3],

ほとんどは、理論、あるいは、自然言語処理の観点での研究であり、本研究で対象とするような知的メカニズムのための連想概念ベースや概念類似度とは異なる。知的判断メカニズムにおける類似度判定の適正さは、最終的には、厳密さより、むしろ人間の感覚的な判断との一致性で左右される。

以下、2章で、まず、我々のねらいとする知的メカニズムの実現に必要な連想概念ベースの構造と、それを用いた連想メカニズムについて述べる。3章以下で、本稿の主題である概念類似度の定量化の問題を定式化し、概念の2次属性までの論理関係を考慮する新方式の提案とその評価について論じる。ここでは、これまでの研究で構築した「基本概念ベース」、すなわち、複数の国語辞書等の語義文から自立語の出現頻度に基づいて属性とその重みを獲得し、更に、その自己参照による新たな属性の追加、及び不要な属性の統計的な除去からなる精練を行うことによって、完全に機械構築した約4万語の概念ベースを前提としている。[4]

2. 連想概念ベースと連想メカニズム

2.1 連想概念ベース

連想概念ベース（約4万語）は、自動学習（概念の自動的追加）と機械精練の容易性の面からできるだけ単純な構造にすることが必要となる。そのため、基本概念ベースの段階では、概念 X を属性語の a_i とその出現頻度等を考慮した重み w_i の集合で構成しているが、連想概念ベースを構成する各概念の構造は、重み情報のない属性語 a_i の集合のみの最も単純な構成とする。（以降単に概念ベースと呼ぶ）

概念ベースの概念構造

概念 X : $\{a_1, a_2, \dots, a_i, \dots\}$

機械精練が適切に進んだ段階では、概念 X の属性語集合 $\{a_i\}$ は概念 X の意味を十分に表現する適切な個数の属性集合（概念集合）であると仮定する。

今回の評価実験では、基本概念ベースの同義語に対する一致数の分析結果（図1）から各概念の属性数は30以下としている。この妥当性及び概念ベースの機械精練については別の機会に報告する。

2.2 連想メカニズム

連想メカニズムは概念ベースと連動しており、入力された概念語／観点から連想した概念集合（属性語）を出力する。もちろん、観点は省略することもでき、その場合には概念語そのものの観点で連想が行われる。

例。

人力：飛行機／乗り物

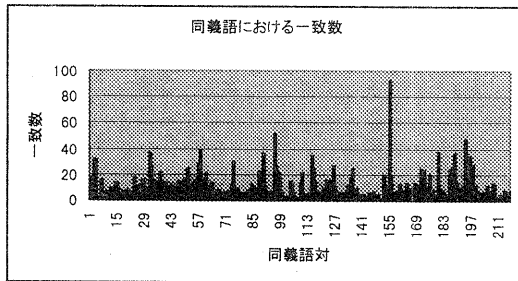


図 1 同義語における一致数

出力：飛行機（推進、噴射、翼、力、回転、出発、音、飛翔、空港、発着、…）

3. 概念類似度の定量化

知的メカニズム構築のための概念類似度定量化問題を定式化する。

3.1 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、そのような単語を概念 A の 1 次属性と呼ぶ。また、概念 A の表記は単語により行う。

$$\begin{aligned} \text{概念 } A &= \text{WORD}_A \\ &= \{\text{WORD}_1, \text{WORD}_2, \dots, \text{WORD}_N\} \\ &= \{a_1^1, a_2^1, \dots, a_N^1\} \end{aligned}$$

概念 A の 1 次属性は単語であるので、その単語を概念と見なせば、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。これらの単語を概念 A の 2 次属性と呼ぶ。

1 次属性 $a_1^1 = \text{WORD}_1 = \{a_1^2, a_2^2, \dots, a_M^2, \dots, a_N^2\}$ さらに、概念の 2 次属性もその意味特徴を表す属性の集合で表現できる。同様に概念 A の n 次属性まで定義可能である。

$$\begin{aligned} n \text{ 次属性 } a_i^n &= \text{WORD}_i \\ &= \{a_1^{n+1}, a_2^{n+1}, \dots, a_i^{n+1}, \dots, a_L^{n+1}\} \end{aligned}$$

したがって、概念 A は n 次までの属性の連鎖で定義されていることになる。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。

3.2 概念の類似度

概念ベースの任意の 2 つの概念間には以下の関係が存在する。

- ・一致：概念名は異なるが概念の 2 次属性までの一致数が極めて高い
- ・類似：概念の 2 次属性までの一致数がかなり高い（一致の度合いを類似度で評価）
- ・無関係：概念の 2 次属性までの一致数が極めて低い

概念の属性として、反意語、相対語、上位分類語、

仲間語、下位分類語、構成要素語、親子語などが入っている場合にはそれなりの考慮が必要となり、連想概念ベースの精練において、これらの関係を処理しておくべきであるが、今回の報告では詳細に立ち入らず、これらも含めすべての関係を類似度として評価している。

3.3 類似度の意味と定義

2 つの概念が似ていることを絶対的な尺度で定義することは困難であり、また、実際問題としても意味がない。従って、次のように、3 つの概念の相对比较で定義する。

p：（概念 X と概念 A）, q：（概念 X と概念 B）の 2 組の概念対を提示された時、人間が感覚として似ていると感じる程度を評価する。具体的には、

p の組が q の組より似ている概念の組みであると判断できる時、概念 A は、概念 B より X に類似していると定義する。この定義は厳密ではないが、人間の会話における意図理解が現実的に、この程度で行われていることから、知的メカニズムの基盤としては妥当と考える。

従って、概念類似度の定量化方式は、2 つの概念 X と概念 A の類似している程度を何らかの数量で表現でき、それを用いて、任意の 2 組みの概念間の類似度を評価した結果が人間の評価結果にできるだけ近い定量化方式が適切な方式であると言うことができる。

3.4 概念連鎖と概念類似度の定量化

我々のこれまでの研究報告[5], [6], [7]も含め、概念類似に関する研究[8], [9]の多くは、定量化のために、概念の属性間の一次独立性を仮定した上で概念ベクトルを構成し、これらのベクトルの内積により類似度を評価している。しかし、この仮定には少なからぬ疑問が残る。また、本研究で基本とする概念処理では、概念間の暗黙の論理関係を積極的に活用することを基本としており、概念間の独立性を仮定することは根底から矛盾することになる。

このため、本研究では、概念のベクトル化はせず、概念の属性集合の類似性を評価する方式をとっている。具体的には、概念連鎖により、属性を n 次属性レベル（今回の報告では 2 次）まで広げ、概念間の論理関係が薄れたところで記号としての一致数を評価することにより、類似度を算出する。

4. 概念類似度の評価モデル

2 つの概念 A と 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。たとえば、「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。「乗り物」という観点で見ると「自動車」の方が「すずめ」より「飛行機」に似ていると考えられるが、「飛ぶ」という観点で見ると「すずめ」の方が「自動車」よりも「飛行機」に似ていると考えるのが普通である。したがって、概念 A と 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。

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る場合
- (2) 観点が指定されていない場合

の 2 つの場合に分けて取り扱う。

また、概念 A と概念 B の類似度評価方式として以下の 3 通りを考える。

<方式 1> 1 次属性中的一致属性数を利用する方式

<方式 2> 2 次属性まで含めた一致属性数を利用する方式

<方式 3> 1 次属性同士の一一致度を利用する方式

方式 1 は、1 次属性のみを利用して類似度を計算する方式である。1 次属性は概念の意味特徴を表す単語の集合であるから、概念 A と概念 B の類似度は、概念 A の 1 次属性と概念 B の 1 次属性がどれほど一致しているかを調べることにより評価できる。しかし、1 次属性のみでは非常に似ている概念同士（同義語）でも単語の一致数はあまり多くなく、精度上の問題がある。図 1 は、同義語対約 200 組について各概念の重要度の高い順に 100 属性を選択し、一致属性数を調べた結果であるが、平均 10% であり極めて低い。そこで、さらに精度を上げようとするのが方式 2 および方式 3 である。

各 1 次属性はそれぞれの 1 次属性（元の概念の 2 次属性）を持っており、単語同士が類似語であればその 1 次属性同士には少なからず同一単語が存在することが予想される。そこで、方式 2 として 2 次属性まで含めて一致属性数を数える方法が考えられる。しかしこの方式では、確かに一致単語数は増加するが、1 次属性と 2 次属性を同等のものとして扱う点に問題がある。2 次属性には 1 以下の重みを掛けることによって評価を下げるという手法も考えられるが、どの程度の重みにすべきか不明である。方式 2 は属性数を元の 1 次属性数の約 2 乗個（同じ属性は除く）にすることにより一致属性数を大幅に増やすことができるが、2 次属性を 1 次属性のレベルに組み入れる事により大きなノイズの混入が懸念される。

そこで方式 3 は、1 次属性同士を比較する際に、単語の完全一致ではなく、その単語が表している概念としての一致度を利用する方法である。すなわち、1 次属性同士の間接一致度としての一致度は、それぞれの 2 次属性同士の一一致単語数から導き、1 次属性同士が単語としては一致していなくても、その一致度合いを見積もれるようにする方法である。

上記 3 通りの案について概念ベースに基づき予備実験を行った結果を図 2 に示す。この実験結果からも 3 方式の中では方式 3 のモデルが類似度の定量化モデルとして最も望ましいと評価できたので、以下では方式 3 のモデルについて詳細検討する。

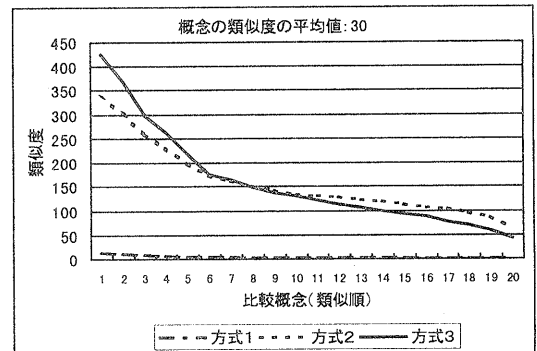


図 2 3 方式の比較

4.1 観点付き類似度評価モデル

いま、2 つの概念 A, B が次式で定義されているものとする。

$$\text{概念 A} = \{a_1^1, a_2^1, \dots, a_l^1, \dots, a_L^1\}$$

$$\text{概念 B} = \{b_1^1, b_2^1, \dots, b_l^1, \dots, b_M^1\}$$

また、観点として概念 C が次式で定義されているものとする。

$$\text{概念 C} = \{c_1^1, c_2^1, \dots, c_k^1, \dots, c_N^1\}$$

このとき、概念 C を観点とする概念 A と概念 B の類似度を $\text{Sim}(A, B | C)$ と書くことにする。

観点 C で概念 A と概念 B を比較するということは、観点 C に関係のある部分だけで概念 A と概念 B を評価するということであり、観点 C に関係のない部分で概念 A と概念 B がいかに似ていようとも関係がない。観点 C に関係がある部分とは、観点 C が持つ 1 次属性に関係がある部分、すなわち、一致度が 0 より大きい部分である。さらには観点 C との一致度の大きさも問題になる。たとえば、観点 C との一致度が 0 ではないが非常に小さい部分での比較と、観点 C との一致度が非常に大きい部分での比較とでは、最終結果に及ぼす影響は後者の方が当然大きくなるべきであろう。ここで、一致度の定義を以下のように与える。

[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、

$$a_i^m = \{a_{i1}^{m+1}, a_{i2}^{m+1}, \dots, a_{iL}^{m+1}\}$$

$$b_j^m = \{b_{j1}^{m+1}, b_{j2}^{m+1}, \dots, b_{jM}^{m+1}\}$$

と表現し、

$a_i^{m+1} = b_j^{m+1}$ (j は $1, 2, \dots, M$ のいずれか) なる a_i^{m+1} の個数を p 個とすると、一致度は次式で定義する。

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

一致度を利用することにより、2つの概念間の類似度はそれぞれの1次属性同士の一一致度の平均として定義できる。ただし、一致度は0から1の実数であるため1次属性同士の対応関係が問題になってくる。いま、ある1次属性 a_i と相手のすべての1次属性 b_j ($j=1 \sim M$) との一一致度を計算したとき、 a_i は一致度が最大の b_j に対応させるべきである。しかし、同じことが他のすべての a_i ($i=1 \sim L$) にも言えるため、問題は複雑になる。これは、1次属性同士を並べるときに、対応する1次属性間の一一致度の合計が最大になるように並べ替える問題である。

このような並べ替え問題は、組み合わせ最適化問題の一種であり、要素数が多くなると組み合わせ爆発を起こすため、真の最適解を求めることはそれほど容易ではない。しかし、文献[11]で提案している「単純法」のように、単純な方法でも比較的最適解に近い値がでることから、本稿で行う実験では並べ替え問題の部分には単純法を利用している。単純法とは、一致度が最大のものを順に選択していく方法であり、その結果が最適解である保証はないが、比較的良好な解が得られるので、ここでの適用には十分であると判断している。なお、より正確に一致度の合計が最大になるように並べ替えたい場合は、遺伝的アルゴリズムなどを用いることができる[11]。

以上の考察より、1次属性同士の一一致度を利用した概念 A と概念 B の観点 C における類似度 $\text{Sim}(A, B|C)$ は以下のアルゴリズム SEAV により評価する。

[観点付き類似度評価アルゴリズム(SEAV)]

1. 概念 C (観点) の1次属性の並びを固定する。
 $C = (c_1^1, c_2^1, \dots, c_k^1, \dots, c_N^1)$
2. 概念 A の各1次属性に対応する概念 C の各1次属性との一致度 $\text{Match}(c_k^1, a_{xk}^1)$ の合計が最大になるように並び替える。このとき、概念 C の各1次属性に対応する概念 A の1次属性は最大1個とする。(概念 C の1次属性の個数 N と概念 A の1次属性の個数 L の大小関係により対

応付けがされない1次属性が残ることになる。)

$$A_x = (a_{x1}^1, a_{x2}^1, \dots, a_{xk}^1, \dots, a_{xN}^1)$$

また、並び A_x の各要素の一一致度を並べたものを、

$$D(A_x) = (d_1^A, d_2^A, \dots, d_k^A, \dots, d_N^A)$$

と書く。ここで、概念 C の1次属性 c_k^1 に対応する概念 A の1次属性がないときには一致度は0とする。

3. 2と同様に、概念 B の各1次属性に対応する概念 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。

$$B_y = (b_{y1}^1, b_{y2}^1, \dots, b_{yk}^1, \dots, b_{yN}^1)$$

$$D(B_y) = (d_1^B, d_2^B, \dots, d_k^B, \dots, d_N^B)$$

4. A_x と B_y の(並びが固定された)各要素間の一一致度 $\text{Match}(a_{xk}^1, b_{yk}^1)$ と、観点 C と A_x の各要素間の一一致度 d_k^A 、および観点 C と B_y の各要素間の一一致度 d_k^B の3者の相乗平均の(相加)平均を概念 A と概念 B の観点 C における類似度とする。すなわち、

$$\text{Sim}(A, B|C) =$$

$$(\sum_{k=1}^N (d_k^A d_k^B \text{Match}(a_{xk}^1, b_{yk}^1))^{1/3})/N$$

4.2 観点なし類似度評価モデル

概念 A と概念 B の観点なし類似度は、観点付き類似度の特別な場合であると考ええる。すなわち、観点がないうことは、比較しようとしている2つの概念以外に第3の概念が陽に指定されていないということであり、あえて観点を指定しようとするれば、(1)概念 A を観点とする場合と、(2)概念 B を観点とする場合の2通りが考えられるが、この2通りは同等であると考えられるので、観点なし類似度は(1)のケースと(2)のケースの平均として計算できる。すなわち、アルゴリズム SAEV で概念 C を A に置き換えた場合と B に置き換えた場合の平均である。以下にアルゴリズムを示す。

[観点なし類似度評価アルゴリズム(SEA)]

1. 概念 A の1次属性の並びを固定する。
 $A_x = (a_1^1, a_2^1, \dots, a_k^1, \dots, a_L^1)$
2. 概念 B の各1次属性に対応する概念 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。
 $B_x = (b_{x1}^1, b_{x2}^1, \dots, b_{xk}^1, \dots, b_{xL}^1)$
3. アルゴリズム SEAV と同様に、概念 A と概念 B との観点 A における類似度 $\text{Sim}(A, B|A)$ は以下になる。
 $\text{Sim}(A, B|A) = (\sum_{k=1}^L (d_k^A d_k^B \text{Match}(a_k^1, b_{xk}^1))^{1/3})/L$

ただし, $d_k^A=1$, $d_k^B=\text{Match}(a_k^1, b_{xk}^1)$ なので, 書き直すと,

$\text{Sim}(A, B|A) = (\sum_{k=1}^L \text{Match}(a_k^1, b_{xk}^1)^{2/3})/L$
あるいは, 1 の場合を相乗平均からはずすことにすれば,

$$\text{Sim}(A, B|A) = (\sum_{k=1}^L \text{Match}(a_k^1, b_{xk}^1))/L$$

4. 同様に, 観点 B における概念 A と概念 B の類似度 $\text{Sim}(A, B|B)$ は,

$$\text{Sim}(A, B|B) = (\sum_{k=1}^M \text{Match}(a_{xk}^1, b_k^1)^{2/3})/M$$

$$\text{Sim}(A, B|B) = (\sum_{k=1}^M \text{Match}(a_{xk}^1, b_k^1))/M$$

5. 概念 A と概念 B の類似度を概念 A の観点からみた場合と, 概念 B の観点からみた場合とに優劣をつけることはできない。したがって, 概念 A と概念 B の (観点なし) 類似度 $\text{Sim}(A, B)$ を A の観点でみた場合の類似度と B の観点でみた場合の類似度の平均とする。すなわち,

$$\text{Sim}(A, B) = (\text{Sim}(A, B|A) + \text{Sim}(A, B|B))/2$$

5. 実験結果と考察

前章までで, 概念間の観点付き類似度の評価モデルと観点なし類似度の評価モデルの提案を行った。本章では, 概念ベースとして文献[10]で作成された $\text{Refine}(K, 1)$ を用いて, 提案した類似度評価モデルの評価実験を行ったので, その結果について述べる。

5.1 観点なし類似度の実験

図3は「指定した概念」と「その概念に比較的類似した概念」の類似度の計算結果の例である。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[10],[12]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見なし, 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。概念ベースの各概念には, その1次属性の重みが出現頻度などを元に設定されているので, 概念ベクトルの各要素の値として, その重みを利用している。それに対して提案方式では, 1次属性の重みは利用していない。

実験結果 (図3) 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), 内積方式ではある程度人間の感覚に近い類似度を出してはいるが, 比較的類似していそうな概念に対する類似度が低く落ち込んでしまっている様子がわかる。それに対し提案方式では, ある程度似ていそうな概念の類似度が比較的高い値にとどまっており, 人間の感覚により近い結果となっている。

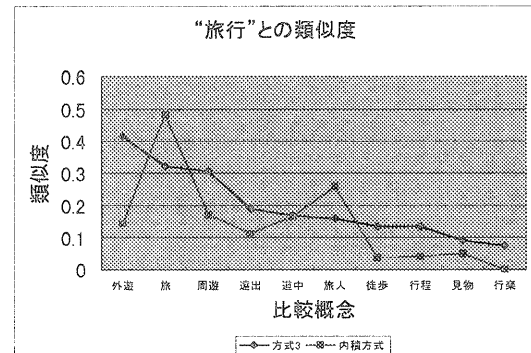
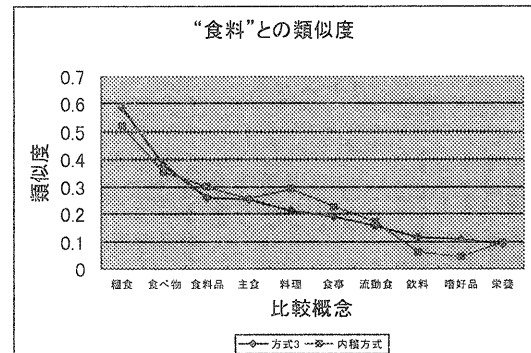
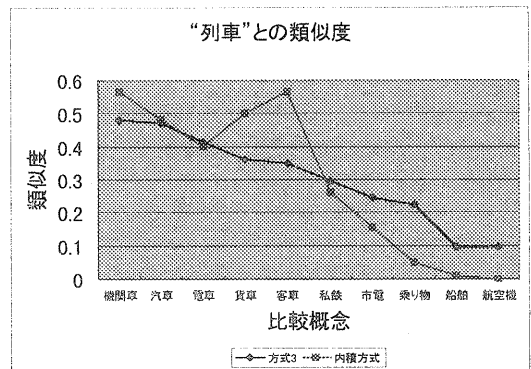


図3 観点なし類似度実験例

5.2 観点付き類似度の実験

3つの概念と2つの観点を使って, 観点付き類似度の評価を行った。たとえば, (豚, 馬, 自動車) という概念と (乗る, 動物) という観点を与えたとき, その各組み合わせに対する観点付き類似度を計算し, 「乗る」の観点で一番大きい類似度となったペアはどれか, 「動物」の観点で一番大きい類似度になったペアはどれかを見て, それが人間の感覚と合っているかを調べた。人間ならば, 「乗る」の観点では (馬, 自動車) が最も類似していると感じるし, 「動物」の観点では (豚, 馬) が最も類似していると感じるであろう。

3つの概念と2つの観点の組を無作為に37通り作成し、実験を行った。表1は人間の感覚と一致した例であり、表2は2通りの内の一方が人間の感覚と一致しなかった例である。2通りとも人間の感覚と一致しなかった例はなかった。37通りの内、成功例は19通りであり、失敗例は18通りであった。約半数が人間の感覚と一致したことになる。

文献[10]では、観点により概念の属性ベクトルを変調する方式により、観点付き類似度を計算している。その方式での成功率は約20%であることが示されているので、本研究による結果はある程度評価できる値であると考えられる。ただし、サンプル数が少ないので、より多くのサンプルでの実験が必要である。

ここで、観点付き類似度の失敗例について検討してみる。表3は表1および表2で利用した概念間の観点なし類似度である。この表から、観点付き類似度が人間の感覚と違う結果となった場合の理由が推察できる。表2の失敗例では、「建物」の観点で(先生,学校)が最高値となっており人間の感覚と合わない。表3の(先生,学校), (先生,警察), (学校,警察)の観点なし類似度を見ると, (先生,学校)の類似度が他の組み合わせの類似度に比べて極端に高い。観点の効果よりも元々の類似度の効果が勝ったために, 「建物」の観点でも(先生,学校)が最高値となったものと思われる。このことは, 他の例を見ても同様である。

表4に「先生」,「学校」,「警察」,「教育」,「建物」それぞれの1次属性(1行目)と2次属性(各1次属性に対して縦にならべたもの)を示す。

6. おわりに

概念類似に関しては, 知識工学や情報工学の立場だけでなく, 認知心理学や, 哲学など幅広い範囲で研究が行われているが多くは基礎理論であり, 実用性をねらったものは機械翻訳, 特定のエキスパートシステムなど目的の限定された範囲の研究である。

本稿では, 汎用的な知的メカニズムの中核機構として概念ベースと概念類似の定量化を捉え, 実用性の面から概念ベースの保守性が極めて重要となることから, できるだけ単純な構造の概念ベースを前提とし, 類似度を定量化する方式について考察した。より適切な類似度定量化のためには, 記号としての一致ではなく, 概念の意味を考慮した定量化方式とすることが有効と考え, 概念のn次属性(今回の実験は2次属性)で類似度を評価する方式を提案した。また提案方式が, 従来提案されている1次属性のベクトル化方式より人間の感覚に合う類似度の定量化方式であることを約4万語の概念ベースを使用した実

表1 観点付き類似度成功例

概念A	概念B	観点X	類似度
豚	馬	乗る	0.040
豚	自動車	乗る	0.029
馬	自動車	乗る	0.056
豚	馬	動物	0.056
豚	自動車	動物	0.019
馬	自動車	動物	0.020

概念A	概念B	観点X	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
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

表2 観点付き類似度失敗例

概念A	概念B	観点X	類似度
先生	学校	教育	0.117
先生	警察	教育	0.037
学校	警察	教育	0.043
先生	学校	建物	0.038
先生	警察	建物	0.010
学校	警察	建物	0.015

概念A	概念B	観点X	類似度
刀	包丁	料理	0.041
刀	俎板	料理	0.028
包丁	俎板	料理	0.169
刀	包丁	刃物	0.066
刀	俎板	刃物	0.037
包丁	俎板	刃物	0.071

表3 観点なし類似度

豚 - 馬	0.134
豚 - 自動車	0.044
馬 - 自動車	0.086
看護婦 - 婦警	0.187
看護婦 - 医者	0.286
婦警 - 医者	0.058
先生 - 学校	0.171
先生 - 警察	0.040
学校 - 警察	0.048
刀 - 包丁	0.070
刀 - 俎板	0.063
包丁 - 俎板	0.222

験により示した。

参考文献

- [1] 岡田 直之：語の概念の表現と蓄積, 電子情報通信学会編, (1991).
- [2] 大須賀 節雄他：知識科学の展開, オーム

社,(1996).

- [3] 鈴木 宏明：類似と思考,共立出版, (1996) .
- [4] 笠原 要,松澤和光,湯川高志,石川 勉,河岡 司：アバウト推論のための多観点概念ベース：構築と評価,人工知能学会全国大会,pp.11-3(1993).
- [5] 笠原 要,松澤和光,石川 勉,河岡 司：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,No3.pp.505-509(1994).
- [6] Kasahara,K.,Ishikawa,T.,Matsuzawa,K.and Kawaoka,T.:Viewpoint-based Measurement of Semantic Similarity between Words, *Proc. 5th International Workshop on Artificial Intelligence and Statistics*,pp.292-302(1995).
- [7] 松澤和光,石川 勉,河岡 司：アバウト推論とその類似性判別機構,AI 学会研究会資料,Vol.SIG-J-9401,pp.103-110(1994).
- [8] 北川高嗣,清水 康,人見洋一：意味の数学モデルとその実現方式について,信学技法,V o 1 .NL95,No.19,pp.1-8(1995).
- [9] 小嶋秀樹,伊藤 昭：意味空間のスケール変換による動的シソーラスの実現,信学技報,Vol.DE93-4,pp.1-8(1995).
- [10] 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,No7.pp.1272-1283(1994)
- [11] 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,情報処理学会春期全国大会,1U-2 (1998) .
- [12] Salton, G. and McGill, M.: *Introduction to modern Information Retrieval*, McGraw-Hill (1983).

表4 概念の1次属性と2次属性

先生(10)										
先生	人	師匠	教師	医師	教員	弁護士	師	教える	指導	
先生	人	先生	教員	医師	教諭	弁護士	先生	知識	教え	
人	他人	教える	教える	医者	学校	資格	僧侶	身	保健婦	
師匠	自分	学問	技芸	聴見	教育	法律	神父	導く	指導	
教師	人柄	芸人	学問	診察	教師	依頼	人	知る	導く	
医師	人数	技芸	人	治療	職員	訴訟	宗教	技芸	進学	
教員	仲人	先師	先生	令	講師	委嘱	軍隊	示	教育	
弁護士	人間	人	宗教	幕府	助教授	登録	専門家	教え	調習	
師	世間	敬称	学校	病気	教授	名簿	旅	手解き	保健所	
教える	世人	芸術	教壇	職業	先生	事件	技能	教授	意図	
指導	人物	教授	教授	江戸	児童	事務	技芸	注意	部分的	

学校(10)										
生徒	学生	児童	中学校	教育	聖学校	学園	本校	小学校	学制	
学校	大学	小学校	学校	人間	学校	学校	学校	初等	制度	
教育	学ぶ	児童	小学校	文教	授ける	私立	分校	学校	学校	
中学校	学問	子供	普通	望ましい	教育	大学	分教場	義務	学区	
受ける	学校	在学	中等	知育	聴力	組織	中心	児童	教育	
音	国学	小学生	三つ	無教育	必要	幼稚園	校	六つ	関する	
児童	学	学齢	施す	教え	耳	施設	対す	学齢	明治	
教え	修学	者	高等	変化	聴覚	多い	本体	施す	制定	
小学校	勉強	学童	旧制	文部省	聴覚	中学	児童	就学	新制	
中等	音	保母	男子	教典	知識	上級	学生	中学校	法令	
新入生	仏教	生徒	五	活動	会話	成る	生徒	学童	近代	

警察(10)										
警察	警察署	国民	生命	都道府県	財産	国家	公安	逮捕	秩序	
警察	警察	公民	生物	府	資産	主権	安寧	逮捕	順序	
警察署	事務	国籍	無生物	事務	財産	国家	公共	現行犯	公序	
国民	都道府県	国家	本源	知事	身上	統治	社会	自由	規則	
生命	役所	統治	存在	行政	身代	人民	安全	令状	正しい	
都道府県	扱う	国	寿命	財	領土	県	安らか	抑留	筋道	
財産	区域	国政	大切	市町村	個人	要素	人々	警察	社会	
国家	管轄	臣民	活動	都	有価証券	邦	無事	被疑者	物事	
公安	留置場	人民	点	総体	総て	鑑	国家	拘束	保安	
逮捕	屯所	同胞	一番	制定	物権	住民	治まる	犯人	関係	
秩序	国民	市民	抽象	広域	土地	国民	保つ	私人	整理	

教育(10)										
人間	文教	望ましい	知育	無教育	教え	変化	文部省	教典	活動	
人柄	教育	願わしい	教育	教育	教える	姿	文化	宗教	前哨戦	
人間	教化	望む	知能	教養	教訓	養れる	学術	教育	休暇	
人間性	学問	変格	知識	修養	教養	活用	教育	本	働く	
人類	文化	形容詞	徳育	学問	教育	権化	官庁	經典	出勤	
人物	行政	動詞	体育	知識	宗教	変わる	中央	教義	防護	
人	関する	欲しがら	高める	望ましい	經典	語形	芸術	教え	拠点	
人格	文部省	感じ	側面	突進	戒戒	進化	学校	経文	休火山	
実存主観	人	活用	豊か	意図的	遺訓	状態	初等	法則	映画	
認識	分野	願う	重要	人間	訓話	性質	外局	説く	発足	
神	文明	教育	認識	知能	事柄	動物	本省	基本	原動力	

建物(10)										
堂	新館	倉	殿堂	楼	建造物	会館	庭園	棟	数寄屋	
建物	建物	建物	建物	建物	塔	集会	建物	棟	茶室	
雅号	新しい	倉	大きい	高い	建造	建物	普請	屋根	建物	
神仏	今迄	家財	堂宇	二階	建物	同郷	橋	建物	茶席	
屋号	別	穀物	神仏	名前	鉄骨	観望	建造物	棟木	障子	
名	会社	倉庫	建造物	料理屋	橋	催し物	工事	刀	水屋	
知らる	本館	所有	軒	下り	船	祭祀	建造物	牛草	美濃紙	
学問	建造物	盗難	物見	国宝	会館	地鎮祭	高い	茶の湯		
母	大きい	所蔵	仏	階	家屋	滑め	建造	数える	貼	
礼	庭園	土蔵	神	楼	卒塔婆	互助	建材	隅	囲う	
祀る	対す	保管	社殿	大きな	倉庫	扶助	大理石	家屋	茶屋	