F-062

重み配分に着目した概念ベースの精練

Refinement of Concept-base Focusing on Weight Distribution

芋野　美紗子‡　吉村　枝里子†　土屋　誠司†　渡部　広一†
Misako Imono　Eriko Yoshimura　Seiji Tsuchiya　Hirokazu Watabe

1. はじめに

近年の情報処理技術の発展は目覚しく，またそれらの技術による各種情報処理システムは私たち人間社会のあらゆる分野に活用されている．高性能化・多機能化した情報処理システムは，社会生活の中でもはや欠かす事のできない存在となっている．しかしそれらシステムの機能向上に伴ってその利用方法も複雑化し，ユーザにとって負担となっている．

このような負担を軽減するためには，ユーザが特別な知識や技能を必要としない利用方法が求められる．その方法の一つとして，人間どうしの会話と同じように自然言語による操作でシステムを利用できればユーザにとって使いやすいシステムになると考えられる．そこで我々は，人間の常識を判断する常識判断メカニズム[1]と，その間をつなぐ未知語処理と想起語処理によるメカニズムを用いて人間の自然な会話や連想をコンピュータで表現することを目指している．

常識判断メカニズムでは，例えば「正月は1月1日」といった時間に関する常識，「赤い」「熱い」といった人間の感覚に関する常識などの意味理解を行うことができる．これら各種常識判断はそれぞれの常識に最低限必要な知識を知識ベースとして所有している．しかし限られた知識だけでは人間の常識すべてを表現することはできない．そこでこの常識判断メカニズムを支えているのが想起語処理，未知語処理を有する連想メカニズムである．

想起語処理とは与えられた語から，その語と関連が強い語を想起する処理である．また，未知語処理は常識判断メカニズムが知識として所持していない語(未知語)を知識として所持している語(既知語)に置き換える処理である．この二つの処理によって各種常識判断が保持しておく知識を最小限に抑えることができる．そしてこの連想メカニズムにおいて重要なのが，ある単語(概念)の特徴を表す語(属性)，そして属性の重要度を定める重みの集合によって定義された概念ベース[2]と，概念間の関連の強さを定量化する関連度計算方式[3]の二つである．

この概念ベースの問題点として，人間の自然な連想にはそぐわない知識が多く定義されていることが挙げられる．具体的には概念の特徴とは関係のない属性が雑音として混じっていたり，逆に関連の強い属性の重みが軽くなっていたりすることがある．これは概念ベースが複数の国語辞書などから自動的に構築された知識ベースであるために起こる問題である．このような問題により，例えば常識判断メカニズムの未知語処理において，まったく関係のない語と置き換えてしまったり，関連の強い語どうしの関連度が低くなってしまったりする．

人間の自然な知識の表現には，正しい連想が必要である．そのためには概念ベースの問題点を解決する必要がある．そこで本稿では概念ベース精練手法を提案する．

2. 連想メカニズム

連想メカニズムでは概念ベースやシソーラス[4]などの知識を利用し，想起語処理と未知語処理を提供している．以下，概念ベースと関連度計算方式について述べる．

2.1. 概念ベース

概念ベースは電子化された国語辞書や新聞記事などから，ある一定のルールに従って自動的に構築された知識ベースである．見出し語(概念)に対して，その特徴を表す語(属性)および属性の重要さ(重み)の対を複数付与することで構成されている．ある概念 A は m 個の属性 A_i と重み w_i (>0)の対によって次のように表現される．

$$\text{概念} A = \{(A_1, w_1), (A_2, w_2), \cdots, (A_m, w_m)\} \quad (1)$$

ここで，属性 A_i を概念 A の一次属性と呼ぶ．概念ベースの具体的な例を表1に示す．

表1　概念ベースの例

概念	属性
雪	(雪,0.61)(白い,0.30)(下る,0.27)・・・
白い	(雪,0.16)(白地,0.14)(色,0.14)・・・
下る	(低い,0.23)(雪,0.21)(雨,0.20)・・・
…	…

また，これらの属性も概念ベースの中で概念として定義されている．つまり属性 A_i を概念とみなして更に属性を導くことができる．概念 A_i の属性 A_{ij} を元の概念 A の二次属性と呼ぶ．図1に概念ベースの構造を示す．

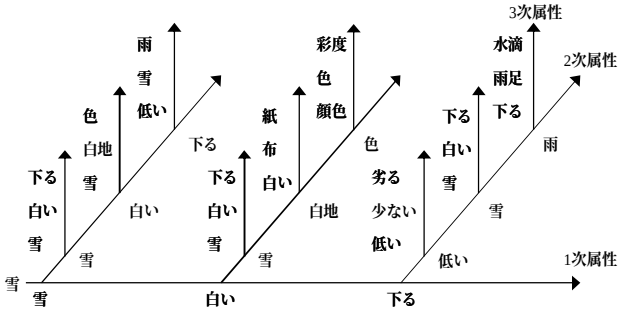

図1　概念ベースの構造

同様に属性を一次，二次，三次，…，N 次と導くことが出来る．つまり概念は任意の次元までの属性連鎖集合により定義されている．

† 同志社大学理工学部
Faculty of Science and Technology, Doshisha University

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

2.2. 関連度計算方式

関連度計算方式とは，概念ベースに定義されている二つの概念間の関連の強さを定量的に表現する手法である．関連度は 0 から 1 の間で値が変動し，概念間の関連が強いほど大きな値を示す．例えば，「林檎」と「蜜柑」の関連度は 0.88，「林檎」と「車」の関連度は 0.0027 となり，同じ果物である概念同士の関連度の方が大きな値を示す．このような定量化手法により，概念間の関連の強さという曖昧なものをコンピュータで処理する事が可能となる．

本稿では，概念間の共通属性を考慮した関連度計算[5]を用いて評価を行う．

3. 評価手法

3.1. 評価データ

精練による概念ベースの精度評価は表 2に示すようなテストデータを用いて行った．任意の基準概念を *X* と置き，この概念 *X* と類義や同義など関連が非常に強い概念 *A*，概念 *A* ほどではないが関連があると思われる概念 *B*，まったく関連のない概念 *C* によって構成されている．この 4 つの概念を一組(*X-A,B,C*)として，人手により人間の常識に沿っていると判断した 500 組を用いて精練後の概念ベースの評価を行った．

表 2　(*X-A,B,C*)評価用データの一部

X	A	B	C
飲食店	食堂	客	得意
飲み物	飲料	液体	選択
病人	患者	治療	磁石
⋮	⋮	⋮	⋮

3.2. 評価方法

概念 *X* と概念 *A* との関連度を *DoA(X,A)*とする．各概念においても同様としたとき，(2)，(3)，(4)の条件式によって評価を行う．

$$DoA(X,A) - DoA(X,B) > AveDoA(X,C) \quad (2)$$

$$DoA(X,B) - DoA(X,C) > AveDoA(X,C) \quad (3)$$

$$AveDoA(X,C) = \sum_{i=1}^{n} DoA(X_i, C_i) / n \quad (4)$$

概念 *X* と関連がない概念 *C* との関連度 *DoA(X,C)*は本来 0.0 となるのが理想である．しかし関連度計算方式の特性上，概念 *X* と概念 *C* に一つでも共通した属性が存在すれば微小な値が算出されてしまう．そこで概念 *C* との関連度 *DoA(X,C)*を誤差とみなし，その平均 *AveDoA (X,C)*をテストデータ全体での平均誤差とする．そして *DoA(X,A)*，*DoA(X,B)*，*DoA(X,C)*それぞれの関連度から平均誤差以上の有意差を得ることができれば正解と見なす．

この評価を全ての組に対して行った上で，正解となったテストデータの組の比率を *C* 平均順序正解率とし，これを概念ベースの精度とする．

なお式(4)における *n* はテストセット数であり，本稿では 500 セット，すなわち *n*=500 となっている．

4. 概念ベースの精練

2.1 節で述べたように概念ベースは機械的に構築された知識ベースである．そのため人間の常識に沿っていない雑音的な属性や，概念と属性の関連にそぐわない重みなども多く付与されている．そこで，新しい概念や属性の追加，適切でない属性の削除や重みの削減，逆に関連の強い属性の重みを増やすなどの操作を行うことで概念ベースの精度を向上させることが必要である．このような操作を概念ベースの精練と呼び，本稿では各概念の属性の重みを一定のルールに従って機械的に変更する精練手法を提案する．

4.1. *N* 次出現回数による精練

ある概念 *X* を考えたとき，*X* は他の様々な概念の中で属性として付与されている．このとき *X* が属性として出現する回数が少なかった場合，その概念 *X* は概念ベース全体の中で重要でない語であると考えられる．そこで属性としての出現回数に閾値を設定し，出現回数が閾値以下の場合には全概念における属性 *X* の重みを小さくする．このような処理を行い，重要でない属性を重み下位にすることで概念ベースの精度向上が得られると考えられる．

概念 *X* の持つ属性が一次属性，一次属性のそれぞれを概念と見て展開した属性が概念 *X* の二次属性，そこからさらに展開したものが概念 *X* の三次属性となる．このように *N* 次まで属性を展開したときに，*X* の出現した回数を *N* 次出現回数と呼ぶこととする．

N 次出現回数が閾値以下だった場合には属性 *X* の重みを定数倍して小さくする．なお，本稿では重みを変更するためにかける定数を重み倍率と呼ぶ．

まず一次出現回数・二次出現回数・三次出現回数それぞれに閾値を定めて精練を行った結果を図 2に示す．なお重み倍率は実験的に検証し，0.1 とした．

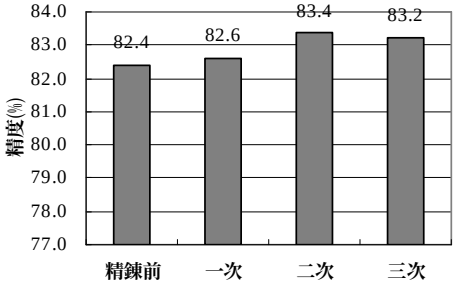

図 2　N 次出現回数のみでの精練結果

結果として全ての場合において，精練前に比べて精度向上が見られた．なお閾値は一次出現回数が 20 回，二次出現回数が 200 回，三次出現回数が 10000 回の時に図 2に示す最も良い精度となった．

向上率を見ると，一次出現回数で 0.2%，二次出現回数で 1.0%，三次出現回数で 0.8%となっており，一次出現回数での向上率を二次出現回数，三次出現回数ともに超えている．この理由として，まず一次出現回数のみを閾値とした場合に重みを削られる属性数が多すぎるという点がある．一次出現回数の閾値が 20 のとき，精練される概念の数は 53230 個となっており，これは概念ベースに定義されている概念の約 45%に達している．もう一つの理由として *N* 次出現回数で閾値を設定すると，概念ベースの連鎖的な構造を考慮する事が出来るという点が挙げられる．二次・三次と概念ベースをさかのぼっても属性としての出現回数が少ない概念の重要度を低いと考え，精練することが可能である．これらのことより概念ベースの

連鎖的構造を考慮にいれた N 次出現回数による精練は有効であることが分かる．

次に，一次出現回数と二次・三次出現回数を組み合わせた精練を行った．つまり精練する属性の選別に，一次出現回数が M 回以下かつ N 次出現回数が L 回以下という二重の条件を与える．これにより概念ベース全体での出現回数に加え，N 次の連鎖構造への影響度合いも配慮することができると考えられる．図 3に精練結果を示す．

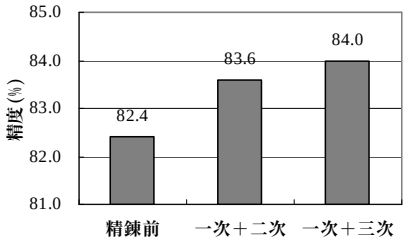

図 3　N次出現回数による精練結果

重み倍率を 0.1 から 0.5 まで 0.05 刻みで変化させて検証を行った．その結果，重み倍率 0.15 の時が一次出現回数と二次出現回数，一次出現回数と三次出現回数の組み合わせ共に最も良い精度となった．一次出現回数と三次出現回数の組み合わせで精度向上率 1.6%と最も大きい成果を得ることができた．

4.2.　二次属性による精練

概念ベースに定義されている概念は，その概念自身に関連の深い概念を属性として持っている．ここで，ある概念 X とその属性 A について考える．属性 A が概念 X の重要な意味定義をなしているならば，逆に概念 A の意味定義をなす属性群の重み上位に概念 X が存在するはずである．以上より，ある概念 X において，概念 X の属性 A から展開した二次属性の重み上位に概念 X 自身が存在しているならば属性 A を重要な属性とみなし，重み倍率をかけて属性 A の重みを大きくする．以上の処理を行い，重要な属性の重みを大きくすることで精度向上を得ることができると考えられる．

属性 A の概念 X における元の重み順位を考慮する場合としない場合で検証を行った．具体的には，属性 A が概念 X において重み上位 30 位の場合とそれ以下の場合で重み倍率を変える処理と，順位に関係なく重み倍率を一律にした処理の二つを行った．まず重み倍率を一律とした場合の精練結果を図 4に示す．

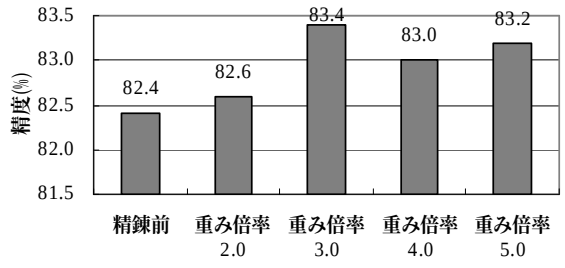

図 4　重み倍率を一律とした場合の精練結果

重み倍率 3.0 の時に 1.0%と最も良い精度向上が見られた．しかし，重み倍率を一律にしてしまうと属性の入れ替わりが起こらないという傾向も見られた．これは重み倍率が一律であるために「関連が強く，元々の重みも大きい属性の重み」と「関連は強いが，元々の重みが小さい属性の重み」が同じように増えることとなり，結果として元々の重みの小さい属性が重み上位に上がってこられなくなっているためである．

関連の強い属性は，本来なら大きい重みを有していなければならない．つまり「関連は強いが，元々の重みが小さい属性の重み」というのは不適当な重み付けになっているということである．このような属性の重みを大きくすることで更なる精度向上が見込めるのではないかと考えられる．そこで倍率を一律にするのではなく，元々の重みを考慮した倍率とすることで精度向上が見込めないかと考え，検証を行った．

まず重みを増やす属性は概念 X の属性 A から展開した二次属性の重み上位 30 個以内に概念 X 自身が存在する場合とする．ここで，属性 A が概念 X の一次属性群の中で重み上位であったか否かを判断し，概念 X の一次属性において重みが小さい属性については重み倍率を大きくする．これにより「関連は強いが，元々の重みが小さい属性」の重みを大きくすることができる．精練結果を図 5に示す．

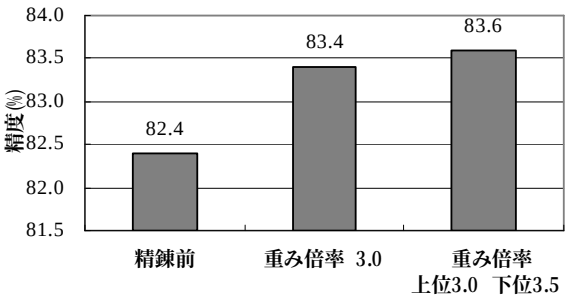

図 5　元の重みを考慮した場合の精練結果

結果として元の重みを考慮した処理では，重み上位 30 位以下の属性について倍率 3.5 とした場合に，重み倍率を一律 3.0 とした場合に比べて 0.2%の精度向上を得る事が出来た．

4.3.　統合結果

以上のように二つの精練手法を提案し，両手法において概念ベースの精度向上がみられた．よってこれら二つの手法を統合して本稿の目的である概念ベースの精練とする．表 3に各精練による関連度平均の変化と精度を示す．なお各種精練手法の順番は以下のとおりである．

①「二次属性」→「N次出現回数」
②「N次出現回数」→「二次属性」

表 3　各精練結果

精練手法	精練前	①	②
AveDoA (*X*,*A*)	0.4491	0.4000 (10.9)	0.4001 (10.9)
AveDoA (*X*,*B*)	0.0988	0.0709 (28.2)	0.0710 (28.1)
AveDoA (*X*,*C*)	0.0030	0.0015 (50.0)	0.0015 (50.0)
C 平均順序正解率	82.4%	83.6%	83.6%

※カッコ内は精練前と比べたときの変化率(%)を示す

最終的な結果として *C* 平均順序正解率は 83.6%となり，精練前と比べて 1.2%の精度向上を得ることができた．

精練前と比べると，*AveDoA* (*X*,*A*)， *AveDoA* (*X*,*B*)，*AveDoA* (*X*,*C*)全ての値が下がっている．しかし精練前と精練後の値の低下を割合でみると *AveDoA* (*X*,*C*)が 50%低下しているのに対して *AveDoA* (*X*,*B*)は約 20%，*AveDoA* (*X*,*A*)は約 10%となっている．3 章で述べたとおり，この評価方法では概念 *X* と概念 *A* が最も関連深く，概念 *X* と概念 *C* は理想上 0.0 となるのが望ましい．よって値の下がる割合が *AveDoA* (*X*,*C*)で最も大きく，*AveDoA* (*X*,*B*)，*AveDoA* (*X*,*A*)と順次小さくなっている本手法の傾向が精度向上につながっているといえる．

5. 考察

提案した各精練手法の傾向および，*C* 平均順序正解率の向上した要因について考察を行う．表 4に各精練手法での関連度の平均および精度を示す．

表 4　各精練手法における関連度の平均

精練手法	精練前	*N* 次出現回数	二次属性
AveDoA (*X*,*A*)	0.4491	0.4567 (1.7)	0.4012 (10.6)
AveDoA (*X*,*B*)	0.0988	0.1030 (4.2)	0.0715 (27.6)
AveDoA (*X*,*C*)	0.0030	0.0033 (10.0)	0.0015 (50.0)
C 平均順序正解率	82.4%	84.0%	83.6%

※カッコ内は精練前と比べたときの変化率(%)を示す

N 次出現回数による精練では，関連度が全体的に増加している．この精練手法では概念ベース全体での出現回数が少ない語の重みを小さくしている．つまり多くの概念で出現している属性の重みが相対的に大きくなり，結果として概念同士の繋がりを多く作り出している属性を重み上位に押し上げている．そのような属性の重みが大きくなることで，ある概念 *X* と概念 *A* の間で共通している属性の重要度が増し，関連度の値が大きくなる．以上のような作用から概念ベース全体での精度向上が得られたと考えられる．

二次属性による精練方法では精練後の関連度が全体的に小さくなっており，*AveDoA* (*X*,*C*)が 50%，*AveDoA* (*X*,*B*)が約 28%，*AveDoA* (*X*,*A*)が約 11%の低下率となっている．4.3 節で述べたとおり，本稿での評価方法では *AveDoA* (*X*,*C*)の値が下がると *C* 平均順序正解率が向上しやすくなる．しかし *AveDoA (X,A)*および *AveDoA (X,B)*の値も低下しているということは，関連の強い概念どうしの関連度の値が下がってしまっているという事である．二次属性による精練は重要な属性の重みを大きくする精練手法であることから，本来であれば関連の強い概念どうしの関連度は大きくなるべきである．理論に反した原因として，二次属性による精練処理に該当しなかった属性にも，概念にとって重要な意味定義をなすものが含まれていたということが挙げられる．つまり，ある概念 *X* にとって重要な属性 *A* から展開した二次属性群に，*X* が含まれていなかったり，重み上位 30 個に入っていない場合が考えられる．

6. 課題

今後の課題としては，手法を組み合わせた場合の関連度の低下が挙げられる． *C* 平均順序正解率は全ての手法において向上しているが，*AveDoA* (*X*,*A*)および *AveDoA* (*X*,*B*)の値は全体的に下がっている．*AveDoA* (*X*,*C*)の低下は維持しつつ，関連の強い概念同士の関連を強くできるような属性操作が必要となる．

また，二次属性による精練手法において，概念 *X* にとって重要な属性であるのに，二次属性の重み上位に概念 *X* 自身が存在していないという問題もあり，このような属性の重みも適切な大きさに変化させる事が出来れば精度向上が見込めるのではないかと考えられる．

7. おわりに

本稿では，概念ベースをより人間の常識的な知識にそったものにするための精練手法を提案した．二次属性による精練は重要な属性を重み上位に，*N* 次出現回数による精練では重要でない属性を重み下位にすると事を目的に精練を行った．そしてこれらの精練の方向性を両方考慮たものを概念ベース全体の精練と考え，最終的な精度はこの二つの手法を用いたものとした．その結果，提案した精練手法を用いることで概念ベースの精度を 1.2%向上させることができた．

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] 土屋誠司, 小島一秀, 渡部広一, 河岡司，“常識的判断システムにおける未知語処理方式”，人工知能学会論文誌,Vol. 17, No. 6, pp. 667-675（2002）．

[2] 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）．

[3] 井筒大志，渡部広一，河岡司，“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”，情報科学技術フォーラム FIT2002，pp.159–160（2002）．

[4] 池原悟，宮崎正弘，白井諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林良彦(編)，“日本語語彙大系”，岩波書店（1997）．

[5] 荒木孝允，渡部広一，河岡司，“共通・類似属性を考慮した概念間関連度計算方式”， 情報処理学会第 68 回全国大会講演論文集，4N-2（2006）．