2007－ICS－147 (25)
2007／3／16

概念ベース内の共起情報に着目した概念間関連度計算方式

荻原 寛† 渡部 広一‡ 河岡 司‡
†同志社大学大学院工学研究科 〒610-0394 京都府京田辺市多々羅都谷 1-3
E-mail： †dtf0741@mail4.doshisha.ac.jp, ‡{hwatabe, tkawaoka}@mail.doshisha.ac.jp

人間と自然に会話できる知的なコンピュータの実現には，単語の意味を理解するシステムの構築が必要であると考える．この実現には，ある概念から他の類似の概念ばかりでなく常識的に関連の強い概念を連想する連想メカニズムが不可欠である．

そこで本論文では，単語の意味を定義している概念ベースの共起情報を利用し，概念間の関連性を評価する関連度計算方式について述べる．従来は，概念ベースの共起情報だけに着目して計算を行っていたが，概念ベースの共起情報だけで無く計算対象概念の類似性に着目し，演算領域を拡大する．そして評価実験により，提案手法の有効性を示す．

The Method of Measuring the Degree of Association between Concepts using Coincidence Information in Concepts Base

Hiroshi OGIHARA† Hirokazu WATABE‡ Tsukasa KAWAOKA‡
† Graduate School of Engineering, Doshisha University
1-3 Miyakodani Tatara Kyotanabe-shi, Kyoto, 610-0394 Japan
E-mail： †dtf0741@mail4.doshisha.ac.jp, ‡{hwatabe, tkawaoka}@mail.doshisha.ac.jp

This paper describes a Calculation Method of Degree of Association which evaluates the relationship between concepts, using the coincidence information between concepts in Concept-Base. In past study, the method was proposed that the Degree of Association is calculated based on only coincidence information between concepts in Concept-base. In this paper, it focuses not only coincidence information but also the similarity of the concepts, and it expands the operation area for calculation. Given this factor, the proposal method extracts some similar concepts to the concept for calculation. This method uses synonym dictionary and a Degree of Match between a concept and its attributes. It is shown that the proposed method is effective by the evaluation experiment.

1 はじめに

知的な処理を行えるコンピュータの実現には，コンピュータにも人間と同じような常識的な判断が行える必要があると考える．そのためには，ある単語から概念を想起し，さらに，その概念に関係のある様々な概念を連想する機能が必要となる．本論文では，そのための仕組みとして概念と概念の関連の強さを定量化する手法を提案する．

従来の共起関連度計算方式[1]は概念ベースの共起情報のみに着目した手法である．そこで，本論文で新たに提案する共起関連度計算方式は，共起情報だけでなく概念の類似性も考慮する．そして，新しく提案した手法が有効的であるかを，評価実験を通して示す．

2 概念ベース

概念ベース[2),3)]とは，複数の電子化された辞書や新聞などから機械的に構築した大規模データベースであり，約9万語の概念が登録されており，各概念は，その概念と関係する語（属性）と，その概念に対する属性の重要度を表す重みの対の集合で定義されている．属性数は概念ごとに異なるが，1概念あたりの平均属性数は約29個である．また，重みは情報量や概念間規則を用いて0～1の実数値で与えられている．概念 A は，その属性と重みと属性数を，それぞれ a_i, w_i, N とした時，以下のように表される．

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

属性 a_i を概念 A の1次属性と呼ぶ．また，属性 a_i も概念であるので，a_i からも同様に属性を導ける．a_i の属性 a_{ij} を概念 A の2次属性と呼ぶ．このように概念から高次の属性を展開できる．

3 概念間の関連性評価方法

3.1 共起情報に基づく関連性評価方法

人が概念間の関連性を判断する場合，「自動車－車」，「電車－汽車」といったように意味的に近いかが重要な判断基準となると考えられる．そこでこれまで概念の意味属性の一致度と重みを利用する意味関連度計算方式[1]により，意味的にどれだけ近いかを判断することで概念間の関連性を評価してきた．ところが意味関連度計算方式では「道－車」，「買う－金」のように連想により導き出せるような語の間の関連性の判断が十分に行えないことが実験によりわかった．人がこれらの語の間に関連があると判断す

るのは，共に使われる頻度が他の語と使われる頻度よりも高いためであると考えられる．そのため意味的な近さを判断するだけの意味関連度計算方式ではこのような語の間の関連性の判断は難しい．

そこで概念ベースの共起情報から語と語の関連性を評価する手法として，共起関連度計算方式が提案されている．共起関連度計算方式には，意味的共起関連度計算方式[1)]と表記的共起関連度計算方式[1)]がある．次節からそれら二つの手法を述べる．

3.2 意味的共起関連度計算方式

概念 A と概念 B の関連が強いほどそれぞれの概念の意味特徴を表す属性集合内に対象とする語が数多く出現すると考えられる．このことから関連性を判断する方法が意味的共起関連度計算方式である（図 1）．

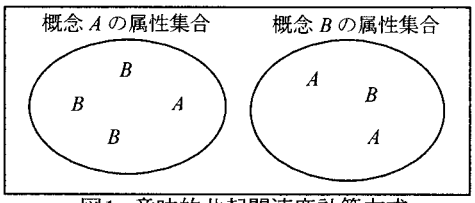

図1 意味的共起関連度計算方式

$$CoCalcN(A,B)=\left(\frac{b_N}{AznumN}+\frac{a_N}{BznumN}\right)\Big/2$$

$AznumN$：概念 A の N 次属性数
$BznumN$：概念 B の N 次属性数
a_N：概念 A が概念 B の N 次属性内に出現する回数
b_N：概念 B が概念 A の N 次属性内に出現する回数

概念ごとに属性内に出現する回数が異なるので，ただ単に出現回数で評価する上記の評価方法では多少問題がある．そこで，情報検索の分野でよく用いられる稀に出現する語を重要とする idf 値[4)]を利用して評価する方法を，以下の式で定義する．

$$CoCalcN\_idf(A,B)=\left(\frac{b_N\times idf(B)}{AznumN}+\frac{a_N\times idf(A)}{BznumN}\right)\Big/2$$

$$idf(t)=\log\frac{N_{All}}{df(t)}+1$$

N_{All}：概念総数 87242
$df(t)$：概念 t が 3 次属性内に出現する概念数

3.3 表記的共起関連度計算

語の対が共出現する頻度から関連性を判断する共起関連度を表記的共起関連度計算方式である（図 2）．以下の式で定義する．

$$Co(A,B)=\left(\frac{df(A\cap B)}{df(A)}+\frac{df(A\cap B)}{df(B)}\right)\Big/2$$

$df(A\cap B)$：概念 A，B が共に一次属性に出現する概念数
$df(A)$：概念 A が一次属性に出現する概念数
$df(B)$：概念 B が一次属性に出現する概念数

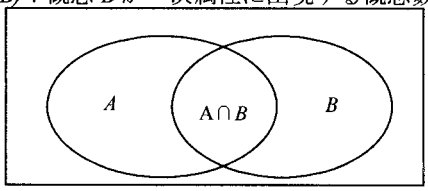

図 2 表記的共起関連度計算方式

4 概念間の共起情報に基づく関連性評価方法

本研究では，人間が常識的にイメージする概念間の距離が近いほど，その関連度計算手法は優れていると判断する．そこで，関連度計算手法の評価をするために，人手によって作成した評価用データを用いて行う．そのために，新しく共起用 *X-ABC* テストデータを構築した．X に対して A は「用意に連想できる」，B は「連想可能」，C は「関連性なし」と定義する．また，yahoo の WEB 検索に *X-A*, *X-B*, *X-C* と AND 検索し，検索 HIT 数が *X-A*＞*X-B*＞*X-C* となったセット用いる．また，WEB 検索 HIT 数の条件を満たしたものを 6 人の被験者に目視評価させた．目視評価の例を図 3 を用いて説明する．図3では，X(秋) に対して，5 人の被験者が「紅葉」の方が連想しやすいと判断したため，「X:秋，A:紅葉，B，さつまいも」とした．そして X(工場) に対しては，「煙突」の方が連想しやすいと判断したため「X:工場，A:煙突:作業」とした．X(子供) に対しては，6 人の意見が分かれたのでテストデータから削除した．

X	候補語1	候補語2
秋	紅葉	さつまいも
工場	作業	煙突
子供	成長	遊具

⇩

X	A	B
秋	紅葉	さつまいも
工場	煙突	作業

図 3 目視評価の例

C には X に対して無関係な語を入れた．このような過程を経て，テストセット（102 セット）を構築した．評価セットの一部を表 1 に示す．

表 1 共起用 *X-ABC* テストデータ（一部）

X	A	B	C
秋	紅葉	さつまいも	従属
野球	投手	審判	縮図
遭難	登山	非常	業界

評価方法は概念 X と概念 A の関連度 $CoCalc(X,A)$，概念 X と概念 B との関連度 $CoCalc(X,B)$，概念 X と

概念 C との関連度 $CoCalc(X,C)$ が，

$CoCalc(X,A) > CoCalc(X,B) > CoCalc(X,C)$

を満たすときを正解とする「順序正解率」で評価を行う．

5　語の類似性を考慮した意味的共起関連度

意味的共起関連度計算方式を，語の類似性に着目して拡張した．拡張方法は，「同義語辞書」，「一致度を用いた拡張」である．

5.1　同義語辞書を用いた拡張

概念 A と概念 B に意味的に類似している語を考慮していく方法として，同義語辞書を用いて行う方式を提案する．同義語辞書とは表 2 のように，ある概念に対して同義とみなされる語がセットとなっているデータベース(32754 セット)のことである．この同義語辞書を用いて意味的共起関連度の拡張手法を述べる．

表 2　同義語辞書(一部)

見出し語	同義語
お父さん	父
胃	胃袋
稲妻	雷光

従来手法の意味的共起関連度の計算の流れは図 4 のようになる．そこに，同義語辞書を用いて対象概念の同義語も，意味的共起関連度の計算対象として扱うように拡張する．拡張した意味的共起関連度計算方式の流れは図 5 のようになる．「n 次属性の中に，対象概念がいくつあるか？」というところまでは同じであるが，その後に，「お互いの属性の中に対象概念の同義語がいくつ存在しているか？」という手順を追加した．

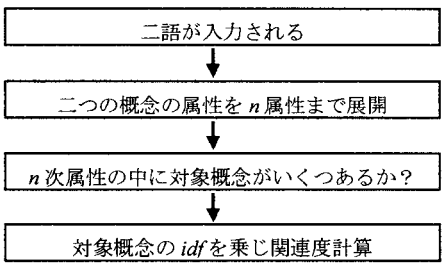

図 4　意味味的共起関連度計算(拡張前)

例を挙げて述べると，「買う－金」を，意味的共起関連度を用いて関連度計算をする場合，従来どおり「金」の n 次属性内に「買う」が何回出現するかを調べる．その後，買うという同義語が同義語辞書内に存在しているか調べる．仮に，同義語辞書内に，「買う－購入」というセットが存在しているとする．その場合，「購入」は「買う」の同義語であるとみなし，今度は「金」の n 次属性内に何回「購入」が存在しているかを調べる．このようにして，「買う」の同義語が同義語辞書内に存在している限り，同じ手順を繰り返して「買う」の n 次属性内に同義語が何回存在しているか調べる．そして，対象概念「買う」の出現回数に idf を掛けた値＋同義語「購入」の出現回数に同義語の idf を掛けた値を求めて，意味的共起関連度計算を行う．

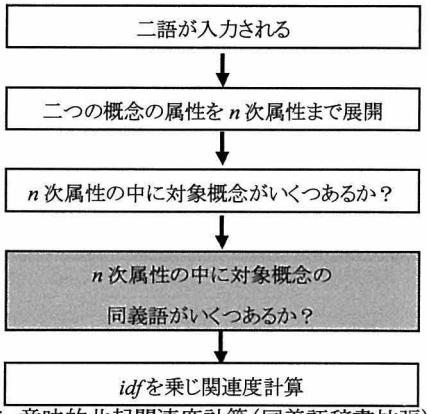

図 5　意味的共起関連度計算(同義語辞書拡張)

5.2　一致度計算を用いた拡張

前節では，定義的な同義語辞書による意味的共起関連度の拡張を提案した．しかし，概念ベースには 9 万の概念が定義されており，必ずしも同義語辞書から同義語を取得できるとは限らない．そこで，対象概念と意味が類似している語を選別する際に，一致度を用いて選出する方法を提案する．

5.2.1　一致度

一致度とは，概念ベースに定義される概念 A, B に対しその 1 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($L \leqq M$)とすると

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

と表現する．概念 A, B の一致度 $MatchWR(A,B)$ は

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

$$\min(\alpha,\beta)=\begin{cases}\alpha & (\beta \geq \alpha)\\ \beta & (\alpha > \beta)\end{cases}$$

(各概念の重みの総和は 1 に正規化する)

と定義する

.

5.2.2　一致度を用いた拡張

一致度を用いた意味的共起関連度計算方式の流れを図 6 に示す．

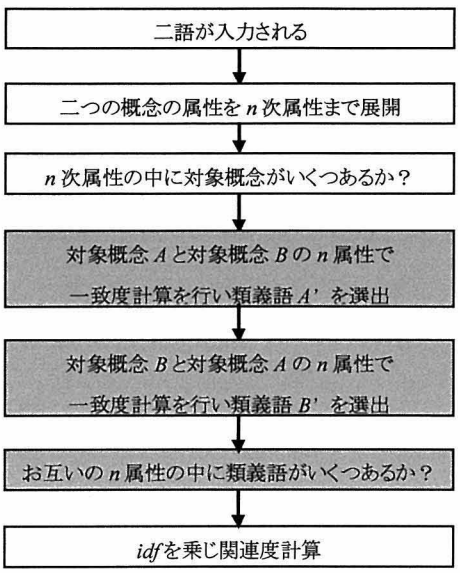

図 6　意味的共起関連度計算（一致度拡張）

5.1 節のときと同様に，「買う－金」を使って説明する．まず，従来の手法どおりお互いの n 次属性の中に対象概念がいくつあるかをカウントする．その後，「買う」と「金の n 次属性」とで一致度計算を行う．このとき，あらかじめ一致度には閾値を与えておき，一致度が閾値以上場合の属性のみを類義語としてみなす．仮に，選出された類義語が「買い入れる」であった場合，「買い入れる」が金の n 次属性内にいくつ存在するか調べる．そして最後に「買う」の出現回数に *idf* を乗じた値＋類義語「買い入れる」の出現回数に類義語の *idf* を乗じた値を計算に用いて意味的共起関連度の値を算出する．評価は節で述べる．なお，一致度による拡張を行う場合においては，過去の研究[1]によって検証されている 3 次属性を用いた意味的共起関連度計算方式を前提とすると，一組の意味的共起関連度計算を行うために，数万回の一致度計算が必要となり，計算時間が実用的ではないと判断したため，2 次属性を用いた意味的共起関連度計算方式を前提とする．

5.3　一致度の閾値の検証

一致度を用いた意味的共起関連度計算方式では，類義語であると判定するための一致度の閾値を設定する必要がある．閾値の値が低いと，ほとんど関係の無いような語まで類義語と選出されてしまうが，閾値の値を上げすぎるとあまり影響のない結果となってしまう．そこで，この手法における最適な一致度の閾値の値を調べる必要があると考えられる．評価結果は 6.2 節に述べる．

5.4　一致度を用いた類義語の属性展開

これまでは入力された概念 A と概念 B の属性内でお互いが何回出現するかで意味的共起関連度計算方式を行ってきたが，語の意味を本質的に考慮するために，5.2.2 節では，一致度によって求められた類義語も考慮した意味的共起関連度計算方式を新たに提案している．では，計算対象の幅を広げるだけでなく，計算対象を捜査する幅を広げたらどのような結果になるかということに着目して検証した．

5.2.2 節で述べた意味的共起関連度計算方式ではお互いの n 次属性の中から一致度によって類義語を選出し，その個数を入力した対象概念の中から調べる．本節ではその一致度で選出した類義語の属性も 2 次属性まで取得して，類義語の n 次属性内に対象概念とその類義語がいくつ存在するかを調べ，意味的共起関連度の値を算出する手法を提案する．イメージを図 7 に示す．

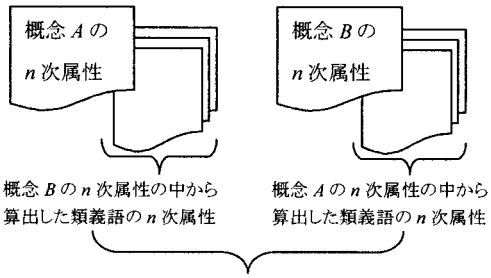

図 7　一致度を用いた類義語の属性展開

それぞれの属性内で求めた「対象概念の出現回数 **idf*＋類義語の出現回数*類義語の *idf*」の値を意味的共起関連度計算方式の通り，それぞれの n 次属性の総数で割る．しかしこのままでは，先ほど求めた値はそれぞれの属性で独立した値となるため，値の合成を行わなければならない．そこで，本論文ではそれぞれの属性内で求めた値の平均値を用いることで値の合成を行った．イメージを図 8 に示す．

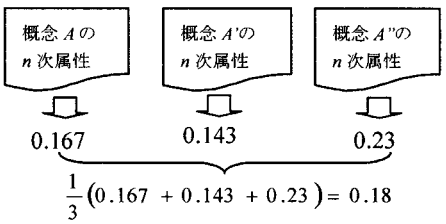

図 8　値の合成の例

そして，概念 A とその類義語の属性を展開して求めた値の平均値と，同様にして概念 B とその類義語

の属性を展開して求めた平均値を用いて，意味的共起関連度計算方式を求める．これが，類義語も n 次属性まで展開して計算する意味的共起関連度計算方式の流れである．評価は 6.3 節に述べる．

6 評価実験

6.1 各種法の評価実験

共起用 *X-ABC* テストデータを用いて評価実験を行った(図 9)．評価方法は順序正解率である．まず，従来手法，同義語辞書を用いた拡張，一致度を用いた拡張の 3 手法で比較する．

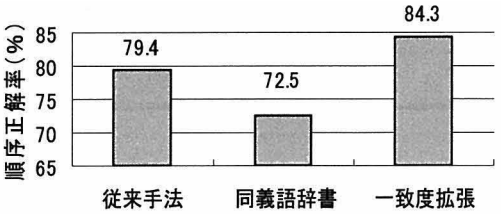

図 9 評価結果（意味的共起）

従来手法では 79.4%，同義語辞書を用いた拡張では 72.5%と従来手法より精度が下がってしまった．順序正解率が低下した原因について調査したところ，*X-ABC* 評価用データ全体としては，*X-A*,*X-B*,*X-C* それぞれにおいて，平均6個ほどの同義語が取得され，同義語辞書による拡張が行われていることが分かった．しかし，ある評価用データでは *X-A* のみに同義語拡張が行われ，*X-B* には行われていない，など，満遍なく拡張が施されていない場合も多数存在していた．このことが順序正解率の精度が低下した原因であると考えられる．しかし，一致度を用いた拡張では精度が 84.3%と，従来手法に比べて約 5%の精度が向上することが出来た．これは，計算対象概念の類義語を，相手の属性内から探すことで，満遍なく拡張の影響が与えることが出来るからである．尚，このときの一致度の閾値は 0.24 とした．

6.2 一致度を用いた拡張の閾値の検証

先の実験で用いた一致度計算の閾値の検証を行う．一致度の閾値を 0~1 まで 0.01 ずつ推移させていった場合に，順序正解率がどのように変化するかを調べ，6.1 節の評価で用いた一致度の閾値 0.24 の閾値が最適であるかを調べる．評価に用いるテストデータは共起用 *X-ABC* テストデータである．検証結果を図 10 に示す．

図 10より，閾値0.24のとき一致度を用いた意味的共起関連度計算方式は最大で 84.3%の正解率を得ることができた．一致度の閾値を下げていくと，類義語に多くの雑音が入り正しく意味的共起関連度が算出できない結果となった．そして閾値をどんどん上げていくと，多少上下はあるが関連度の値が低下していき，一致度の閾値 0.69 からは値が収束して 79.4%となった．

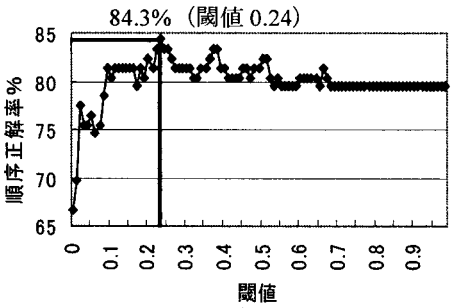

図 10 一致度の閾値の検証

6.3 一致度を用いた類義語の属性展開

一致度を用いた類義語の属性展開の評価結果を示す．尚，対象概念の属性内から類義語を選出する一致度の閾値は6.2節で述べた0.24を用いている．ここで，対象概念の中では 0.24 という閾値が一致度を選出するのに最も良い値となったが，類義語の属性を展開する場合にはこの 0.24 は最適な数字なのかを調べてみた．6.2 節の閾値調査と同じように，展開する類義語の一致度の閾値を0～1まで0.01刻みで調べていった．その結果を図 11 に示す．

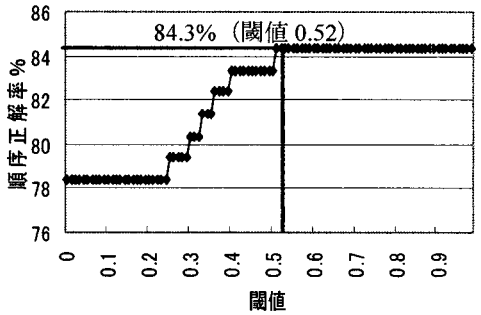

図 11 展開する類義語の閾値の検証

図 11 から，閾値が低い，すなわち展開する類義語が多数存在する場合では，精度は低下している．そして，一致度の閾値を上げて，展開する類義語をしぼっていくと精度は上昇していくが，閾値 0.52 のとき一致度を用いて拡張した意味的共起関連度計算方式の順序正解率の精度 84.3%に収束してしまう．この結果から，現段階では類義語を展開するメリットはないと言える．

順序正解率の精度が上がらない原因として，対象概念と類義語の属性を展開して求めた値の合成に問題があると考えられる．その例を図 12 に示す．図 12 では，対象概念 *A*，類義語 *A'*，*A"*の 2 次属性の中に，対象概念 *B* とその類義語がいくつあるか調べ，それに *idf* を乗じた後にそれぞれの 2 次属性の総数で割った値を示している．この例は，類義語の 2 次属性の中に対象概念 *B* とその類義語 *A'*，*A"*が極端に少なかった場合である．本論文では，値の合成は

それぞれの値の平均値を用いているので，類義語 *A'*, *A"*の値が本来であれば高い関連性を示している対象概念 *A* の値を引っ張っていることになる．この傾向がテストデータの中でも特に *X-A* において顕著に現れ，関連度の値は *X-A*＜*X-B* と逆転してしまう結果が起きていた．このことにより順序正解率が低下したと考えられる．今後，走査領域を類義語まで拡大した方式を用いるとするならば，それぞれの独立した意味的共起関連度の値を適正に合成する手法が必要となると考えられる．

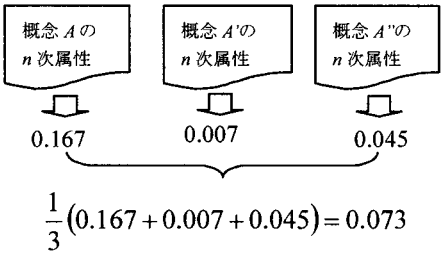

図 12　類義語の属性展開の場合の失敗例

7　語の類似性を考慮した表記的共起関連度

前節までは意味的共起関連度計算方式に対して，同義語辞書や一致度を用いた拡張を行ってきた．本節では，表記的共起関連度計算方式に対して，意味的共起関連度計算方式の拡張における評価結果を基に，同義語辞書ではなく一致度を用いた拡張方式を提案する．

7.1　一致度を用いた拡張

表記的共起関連度計算方式の拡張手法として，計算対象概念の 1 次属性から一致度を基に類義語を導出し，捜査領域を拡張する手法について述べる．具体的には，まず対象概念の1次属性と対象概念とで一致度計算を行う．このときの類義語を選定する一致度の閾値の検証は，5.2 節で述べる．そして，一致度で求めた類義語を 1 次属性に持つ概念の集合(1 次概念)を取得し，対象概念の 1 次概念と類義語の 1 次概念を結合させる．そして，概念 *A*＋類義語 *A'*と概念 *B*＋類義語 *B'*の 1 次概念同士で共通の概念を見つけ出し，共起関連度計算を行う．イメージを図 13 に示す．

図 13　表記的共起関連度計算の拡張

7.2　評価実験

共起用 *X-ABC* テストデータを用いて評価実験を行った(図 14)．評価方法は順序正解率である．

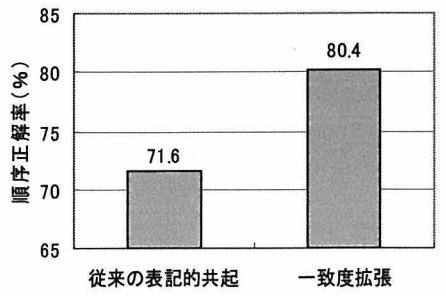

図 14　評価結果(表記的共起)

類似した概念を考慮することにより，順序正解率を一致度の閾値 0.29 の時に 80.4%まで向上させた．尚，このとき用いた一致度の閾値は 0.29 である．この 0.29 という閾値は，6.2 節で用いた実験方法を用いて検証し，求めた値である．

8　おわりに

本論文では，従来の共起関連度計算方式に概念の類似性を考慮すること提案した．そして，従来手法よりも意味的，表記的それぞれの共起関連度計算方式で順序正解率の精度を上げることができた．このことから，共起情報に加え，概念の類似性も考慮したほうが有効的であることがわかった．

謝辞

本研究は文部科学省からの補助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における研究の一環として行ったものである．

参考文献

1) 渡部広一，奥村紀之，河岡司，"概念の意味属性と共起情報を用いた関連度計算方式"，自然言語処理，Vol.3，No.1，pp.53-74，2006.
2) 奥村紀之，北川晋也，渡部広一，河岡　司,"概念ベースの分析と精錬"，同志社大学理工学研究報告，Vol.46，No.3，pp.133-141，2005.
3) 奥村紀之，渡部広一，河岡司，"電子化新聞を用いた概念ベースの拡張と属性重み付与方式"，情報処理学会研究報告，2005-NL-166，pp.55-62，2005.
4) 徳永健伸，"情報検索と言語処理"，東京大学出版会，1999.