F-002 概念ベースにおける属性の共起情報を用いた関連度計算方式の拡張

The Expansion of the Method of Measuring the Degree of Association between Concepts using Attributes of the Concepts and Coincidence Information in Concept-Base

辰己 直彦 † 青田 正宏 † 渡部 広一 † 河岡 司 †
Naohiko TATSUMI Masahiro AOTA Hirokazu WATABE Tsukasa KAWAOKA

1. はじめに

コンピュータ上で知的な処理を実現するためには，人間の持つ常識的な判断能力をコンピュータに与える必要がある．そのためには，ある単語からその語に関連のある様々な概念を連想する機能が必要となる．また，基の語と想起された語との関連性を定量的に評価するための計算手法が必要である．

従来，関連性の評価方式として，概念の特徴を表す属性の一致度と重みを利用する意味関連度計算方式 [1] が用いられてきた．本稿では，従来方式の改良と問題点の解決を目指し，概念と概念の共起情報を用いた共起関連度計算方式の提案，また，共起関連度計算方式と意味関連度計算方式を複合利用する関連性評価方式を提案する．

2. 概念ベース

概念ベースは言葉に関する大規模なデータベースであり，ある概念 A は，その概念の意味特徴を表す属性 a_i と，この属性 a_i が概念 A を表す上でどれだけ重要かを表す重み w_i の対で表現される．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 つの概念であるので，a_i からも同様に属性を導くことができる．このように，概念は属性の n 次元連鎖により定義されている．

概念ベースには，約 9 万の概念があり，1 概念あたりの平均属性数は 30 個である．属性の展開は, 三次以上行うと関連度計算の評価実験の正解率が下がることが知られているので，本稿では属性の展開は三次までとした (関連度計算の評価方法については次節 (3.) 参照).

3. 関連度計算の評価方法

表 1: 評価用データ (一部)

基準概念X	高関連概念A	中関連概念B	無関連概念C
図書館	書物	勉強	気軽い
学習塾	勉強	子供	全納
下着	衣服	肌	舞楽

表 1 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（基準概念）であり，概念 A は概念 X と極めて密な概念，概念 B は概念 X に密な概念，概念 C は概念 X に疎な概念である．すなわち，基準概念 X に対して A が非常に関連が強く，B は関連があり，C はほとんど関連がない概念である．表 1 の各評価用データの組に関して

$$Rel(X, A) - Rel(X, B) > AveRel(X, C)$$
$$Rel(X, B) - Rel(X, C) > AveRel(X, C)$$
$$AveRel(X, C) = \sum_{i=l}^{2370} Rel(X_i, C_i)/2370$$

を満たせば，その関連度計算結果を正解とし，それ以外は不正解とする．ただし，$Rel(a, b)$ は概念 a と概念 b との関連度である．このような評価用データを人手で作成し，作成者以外の 3 人に判定させ，3 人ともが正しいと判断した 2370 組を採用した．2370 組のデータの中で正解したものの割合で関連度計算の精度評価を行う．

4. 意味関連度計算方式

4.1 計算方式

2 つの概念 A，B の意味関連度 $MR(A, B)$ は，概念 A，B の一次属性のすべての組み合わせについて，二次属性を利用し一致度（一致する属性のうち小さい方の重みの和）を求め，一致度の和が最大になるように一次属性の組み合わせを作る．ここで，一次属性が完全一致する場合は別扱いにする．これは概念ベースには約 9 万の概念が存在し，属性が一致することは稀である．従って，属性の一致の扱いを別にすることにより，属性が一致した場合を大きく評価する．具体的には，対応する属性の重みの大きさを，重みの小さい方の値とする．このとき，重みの大きい方は，その値と小さい方の重みの差をとり，再度，他の属性と対応をとる．

組み合わせが決まった後の，概念 A，B の一次属性をそれぞれ，$a_i^{'}$，$b_i^{'}$，その重みを，$u_i^{'}$，$v_i^{'}$，とし，対応の取れた属性の組み合わせが T 個の場合，

$$A = \{(a_1^{'}, u_1^{'}), (a_2^{'}, u_2^{'}), \cdots, (a_T^{'}, u_T^{'})\}$$
$$B = \{(b_1^{'}, v_1^{'}), (b_2^{'}, v_2^{'}), \cdots, (b_T^{'}, v_T^{'})\}$$

となる．

概念 A，B の意味関連度 $MR(A, B)$ は，

$$MR(A, B) = \sum_{i=1}^{T} Match(a_i^{'}, b_i^{'}) \times (u_i^{'} + v_i^{'}) \times \frac{1}{2} \times (min(u_i^{'}, v_i^{'})/max(u_i^{'}, v_i^{'}))$$
$$Match(a_i^{'}, b_i^{'}) = \sum_{a_{ip}^{'} = b_{iq}^{'}} min(u_{ip}^{'}, v_{iq}^{'})$$

（$u_{ip}^{'}$, $v_{iq}^{'}$ は $a_i^{'}$, $b_i^{'}$ の一致する一次属性の重み）

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

となる．意味関連度は対応する一次属性の一致度（$Match$）と，それらの属性の重みの平均，および重みの比に比例すると考える．以後，意味関連度計算方式を MR とする．

4.2 評価実験

評価用データを用いて，意味関連度計算方式の評価実験を行った結果，正解率は 71.1%であった．

この方式では,「買う－金，買う－雨」,「道－車，道－点」など「X-B,X-C」の判定が正しくできていないものが多数見られた．

4.3 考察

概念の特徴を表す属性の一致度合から関連性を判断する従来方式では,「買う－購入」のように近い意味を持つ語の関連性は正しく判定されていた．しかし,「買う－金」のように，連想により導き出せるような語の関連性の判定は困難であったと言える．

人が「買う－金」の方が「買う－雨」よりも関連が強いと判断できるのは,「買う－金」という語の組み合わせの方が,「買う－雨」という語の組み合わせよりも一般的であると判断しているためだと考えられる．そこで，概念間の共起情報を考慮し，関連性を判定するための計算方式が必要であると考えられる．

5. 共起関連度計算方式

5.1 表記的共起関連度計算方式

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以下の式で定義することにする．

$$Co(A,B) = (\frac{df(A\&B)}{df(A)} + \frac{df(A\&B)}{df(B)})/2$$

$df(A\&B)$：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

『表記』の対がどれだけ出現するかを評価するため，この手法を表記的共起関連度計算方式 (Co) と呼ぶことにする．

5.2 意味的共起関連度計算方式

概念 A と概念 B の関連性が強いほど，それぞれの概念の N 次属性集合内に対象とする語が多数存在すると考えられる．言い換えると，概念 A の N 次属性内に概念 B が多数存在すると思われる．このように関連性を判断する方法を意味的共起関連度計算方式 ($CoCalcN$) と定義する．

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

$AznumN$：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，出現回数に依存するため問題がある．そこで，情報検索の分野で広く用いられる idf を利用して評価する方法 ($CoCalcN\_idf$) を，以下の式で定義する．idf は稀に出現する語を重要視する方式である．

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

$idf(t) = \log \frac{N_{All}}{df(t)} + 1$
$df(t)$：概念 t が N 次属性内に出現する概念数
$NAll$：概念総数 87242
(N=1,2,3)

5.3 評価実験

評価用データを用いて共起関連度計算方式の評価実験を行ったところ，図 1 のようになった．

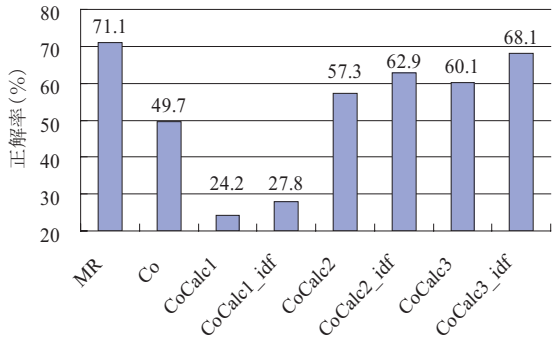

図 1: 共起関連度評価結果

最も精度がよかったのは，三次属性まで展開して idf 値を用いる意味的共起関連度計算方式であり，正解率は 68.1%であったが，従来方式より精度がよくならなかった．そこで，従来方式と共起関連度計算方式を複合利用することにより精度向上を試みた．

6. 意味共起関連度計算方式

6.1 計算方式

従来の意味関連度計算方式と共起関連度計算方式を合成して評価する方式を以下で定義し，この計算方式を意味共起関連度計算方式 (MCR) と呼ぶことにする．

$$MCR(A,B) = \frac{MR(A,B) + C_w \times CR(A,B)}{1 + C_w}$$

$MR(A,B)$：意味関連度　$CR(A,B)$：共起関連度
C_w：共起関連度重み

意味共起関連度計算方式では，共起関連度に重みを付与し，意味関連度に足し合わせ値を算出する．

6.2 評価実験

評価用データを用いて，意味共起関連度計算方式の評価実験を行う．評価方法は，共起関連度に付与する重み C_w を 0.0 から 10.0 まで 0.005 間隔で変化させ，評価用データの正解率が最大になる C_w を調べた（図 2，棒グラフ上記の数値は C_w の値）．

図 2 は，意味関連度計算方式に対して，どの計算方式を組み合わせたときに最もよい精度が得られるかを調べた結果である．図から，最も精度が良かったのは，意味関連度計算方式と共起関連度計算方式 $CoCalc2\_idf$ に重み C_w=8.970 を付与して組み合わせた意味共起関連度計算方式の正解率 76.5%で，従来方式の正解率 71.1%と比べ，5.4%精度が向上したと言える．

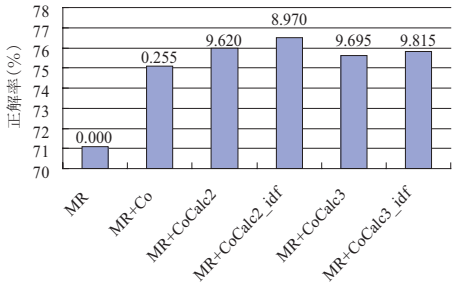

図 2: 意味共起関連度評価結果

次に，意味共起関連度計算方式の評価用データに対する依存性を検証するため，新たな評価用データを作成し，評価実験を行った．評価用データに対し，最高正解率を得るためのパラメータ実験 (6.2) を行ったため，評価用データに特化していないことを実証することで提案方式の有効性を示す．

7. 評価用データ

7.1 評価用データの自動作成

人手で作成した評価用データに対して同義の意味の概念を用いて変換を行っても，基準概念 X に対する概念 A，B，C の人目で判断したときの関連の強さの順は，変わらないと考えることができる．また，異なる概念の特徴を表す属性は，たとえ同義関係の語であったとしても，完全に一致することはないため，評価用データを機械的に拡張することが可能である．

そこで，評価用データ内の概念 X，A，B それぞれに対して，同義語リストを用いて各概念を同義語に変換する．同義語リストとは表 2 に示すような見出し語と同義語から構成されたリストである．同義語リストには，日常会話でよく用いられる単語を対象に作成した手動構築同義語リスト (2030 セット) と，電子化された国語辞書から概念構造情報 [2] を作成し，その中から同義の関係となっているものを選出した自動構築同義語リスト (3 万 5000 セット) がある．これら 2 つの同義語リストを用いて，評価用データを拡張する．

表 2: 同義語リスト (一部)

見出し語	同義語
十二月	師走
ドクター	医者
和室	日本間

概念 C は無関連概念であるから変更を加えない．評価用データに対し無関連であることが保証されているため，同義語置換を行っても関連性に何ら変化を及ぼさないためである．

手動構築同義語リストを用いた場合，2976 セットの評価用データ (データ A) が作成され，自動構築同義語リストを用いた場合，11895 セットの評価用データ (データ B) が作成された．これらの評価用データに対し，それぞれ 400 セットをランダム抽出した (必要最低限の標本サイズは統計学的に求めた)．そして，基準概念 X に対して関連の強さが A,B,C となっているかどうか人目で判定した結果，図 3 のようになった．

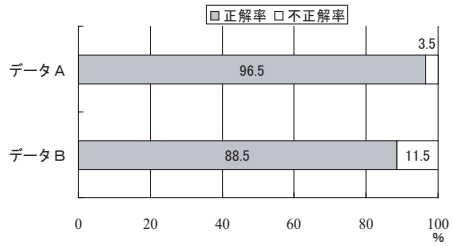

図 3: 目視による評価結果

データ B は 11.5%のデータが基準概念 X に対して関連の強さが A,B,C となっていないので，テストデータとしては利用できないと考えられる．そこで，データ A を評価用データとして用いる．以後，既存の評価用データを mda とし，新たな評価用データを拡張評価用データ $smda$ とする．

7.2 評価実験

評価用データと拡張評価用データを用いて，意味関連度計算方式と意味共起関連度計算方式の評価実験結果を，図 4 に示す．

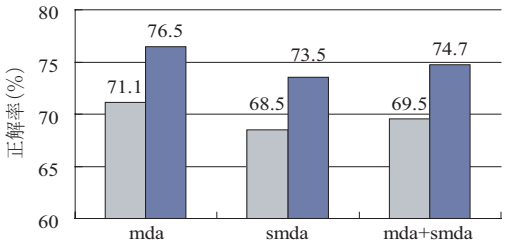

図 4: 各評価用データに対する関連度評価結果

棒グラフの左側が意味関連度計算方式の正解率で，右側が意味共起関連度計算方式の正解率である．評価用データをそれぞれ個別で用いた場合と，両評価用データを組み合わせた場合，どちらの場合にも意味共起関連度計算方式の方がよい結果が得られた．

8. おわりに

今回提案した意味共起関連度計算方式により，従来の意味内容に重点を置いた関連度評価手法よりも的確に関連度の判断ができるようになった．また，評価用データを機械的に増加させる手法を開発し，評価用データを倍増させた実験により，提案方式の有効性を確認した．

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“ 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] 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.

[2] 小島一秀, 渡部広一, 河岡司. 電子化国語辞書を用いた概念ベース自動構築における前処理の自動化. 情報処理学会第 58 回全国大会.