F-003

関連度における共通閾値の存在と応用

Existence and Application of Common Threshold of the Degree of Association

小島 一秀† 渡部 広一‡ 河岡 司‡
Kazuhide Kojima Hirokazu Watabe Tsukasa Kawaoka

1 はじめに

語間の関連度は語間の関係の強さを示す値[1][2]であるが，完全に正しい関連度がどのような値の集合であるかという考察は十分ではない．もし，完全に正しい関連度の特性がわかれば，目標とすべき関連度の特性が明確となり，関連度の評価や関連度計算方式の構築など関連度の関わる多くの場面で有効である．本稿では，関連度が完全に正しいときに関連度の定義のみから共通閾値の存在が導かれることを示すと同時に，共通閾値の応用について述べる．共通閾値とは，全ての語で共通な，関係の強さを識別できる関連度の閾値である．

2 関連度の定義と共通閾値

関連度の定義は次のようになっている．

- 関連度は語間の関係の強さを示す値である．
- 関連度は 0 以上 1 以下で関係が強いほど値が大きい．

語間の関係には観点付き関係など様々なものがあるが，本稿では特に観点などのない語間の単純な関連度を取り上げる．ただし，共通閾値は，単純な関連度に限らず，観点付き関連度や文書間関連度でも，関連度が完全であれば存在する．共通閾値とは次のような値である．複数の強い関係 r_i と複数のそれより弱い関係 s_j があるとき，関連度が完全であれば，ある語を基準に関連度の計算をした場合，ある閾値で強い関係 r_i と弱い関係 s_j を識別できるはずである．なぜなら，関連度が完全なら，全ての r_i の関連度が，全ての s_j の関連度より大きくなるためである．この閾値が関連度を計算する語に関係なく等しいとき，共通閾値と呼ぶ．

3 識別率

仮に完全な関連度において共通閾値が存在しているとすれば，共通閾値で関係の強さをどれだけ正確に識別できるかで，関連度を評価できる．この節では関係の強さを同義と類義の高関連，ある程度の関係のある中関連，全く関係のない無関連の 3 段階とする．評価用のテストデータは，基準語 X，基準語 X に対して高関連の語 A，中関連の語 B，無関連の語 C の 4 語で 1 レコードとなっている（表 1）．

表1 テストデータ

基準語	高関連	中間連	無関連
樹木	木	木の葉	頭
天気	天候	雨	写真

まず，共通閾値を求める．具体的には，テストデータの高関連と中関連の関連度を一つずつ閾値 t_{ab} とし，その閾値以上となった高関連の関連度の個数 n_a とその閾値未満となった中関連の個数 n_b を調べて，

$$(n_a/N_a+n_b/N_b) / 2 \quad (1)$$

が最高になる t_{ab} を探す．ただし，N_a，N_b，N_c はそれぞれテストデータにおける高関連，中関連，無関連のデータ数である．表 1のようなテストデータの場合，テストデータ数 $N_t = N_a = N_b = N_c$ となる．さらに，中関連と無関連のデータの間でも同様の方法で閾値 t_{bc} を求める．求めた閾値を用いて，識別率は，

$$(n_a/N_a+n_b/N_b+n_c/N_c) / 3 \quad (2)$$

となる．ただし，n_b，n_c はそれぞれ，閾値 t_{bc} 以上 t_{ab} 未満となった中関連の関連度の個数と，閾値 t_{bc} 未満となった無関連の関連度の個数である．式(3)のような評価値も考えられるが，次のような理由で式(2)を識別率としている．

$$(n_a+n_b+n_c) / N_t \quad (3)$$

式(2)は各関係の強さを正しく判定できた率の平均であり，各強さの関係のデータ数の違いによる影響が小さい．もし，識別率を式(3)のようにすると，例えば，無関連のデータを多数入力した場合，常に無関連と識別するような，他の関係の識別を無視した閾値が設定される可能性が高い．現実には無関連が多いため，それを多数入力するのは適切なようにも見えるが，確率的に最も高いという理由で常に無関連と答えるようなシステムより，どのような関係の強さにも一様に適切に判断できる方が人間らしい能力としては自然であると考えるためである．

ここでは，3 段階の関係の強さを扱ったが，識別率は 3 段階に限らず 2 段階以上であれば何段階の関係の強さにも対応可能である．

4 共通閾値の存在

全ての語間の関連度を誤りなく計算できる状況では，共通閾値が存在しないと，関連度の定義を満たすことが極めて困難であることを示す．まず，全ての語間の関連度が誤りなく計算できると仮定する．ここでは，考察を単純化するために次のような仮定を行う．

- 高関連と中関連を合わせて有関連とし，関係の強さを有関連と無関連の 2 段階とする．
- 全ての語は，有関連と無関連を識別する閾値が t_1 の低閾値語と，t_2 の高閾値語のどちらかであるとする（$t_1<t_2$）．

†大阪外国語大学情報処理センター，kkojima@indy.doshisha.ac.jp
‡同志社大学大学院工学研究科，{watabe, kawaoka}@indy.doshisha.ac.jp

高閾値語“電車”を基準にして考察する．“電車”とそれに対して有関連の低閾値語“走る”との関連度は，高閾値語の閾値 t_2 以上の領域が適切となる（図 1）．有関連であるとき，低閾値語“走る”の閾値 t_1 以上，t_2 未満の領域は，低閾値語“走る”にとっては適切であるが，高閾値語“電車”にとっては不適切であるため，最終的には不適切な領域となる．

“電車”とそれに対して無関連である低閾値語“書く”との関連度は，t_1 未満の領域が適切である（図 1）．無関連であるとき，t_1 以上，t_2 未満の領域は，高閾値語“電車”にとっては適切であるが，低閾値語“書く”にとっては不適切であるため，最終的には不適切な領域となる．

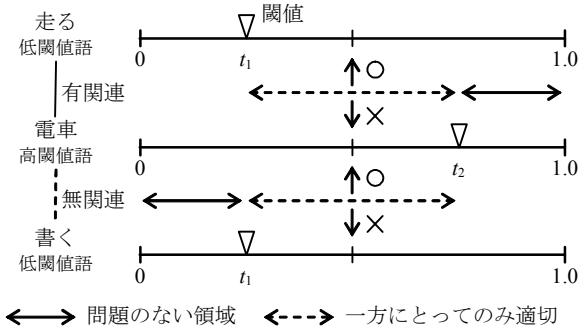

図1 閾値の異なる語間の関連度

以上から，閾値の異なる語同士の関連度計算では，有関連のときは t_2 以上にし，無関連のときは t_1 未満にすることによって，関連度を計算する語の閾値の差の領域に入る値が出ないようにする必要があることがわかる．しかし，関連度は 0 から 1 までの値であり，ある領域を無効にするような計算方式の実現は極めて困難である．そうなると，各語の閾値の差をなくさなければならず，各語の閾値は唯一となりそれが共通閾値となる．また，実際の関連度では 3 段階程度の関係の強さを扱うことも多いが，多段階になればなるほど閾値の差を持たせるのは困難である．

5 共通閾値のない関連度

共通閾値の存在しない状況で完全な関連度が，仮に実現したとしても，共通閾値が結果的に発生する．

関係の強さを有関連，無関連の 2 種類とする．全語数を N とし，語 a_i の閾値を t_i とするが，t_i 以上が有関連である．ただし，$1 \le i \le N$ である．

まず，$N = 4$ の場合で関連度の閾値がどれだけの関連度に制約を与えるかを調べる．4 語の閾値を $t_1 < t_2 < t_3 < t_4$ とし，4 語の組み合わせから導かれる関連度に対して制約を与える閾値を，有関連，無関連のときに分けて表 2に示す．

表2 関連度を制約する閾値

	有関連				無関連			
	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
a_1	t_1	t_2	t_3	t_4	t_1	t_1	t_1	t_1
a_2	t_2	t_2	t_3	t_4	t_1	t_2	t_2	t_2
a_3	t_3	t_3	t_3	t_4	t_1	t_2	t_3	t_3
a_4	t_4	t_4	t_4	t_4	t_1	t_2	t_3	t_4

全ての関係が有関連である場合，閾値 t_i が制約を与える関連度の数は，

$$2i - 1 \tag{4}$$

となる．有関連の場合は大きい方の閾値が有効であるため，表 2では各閾値が L 字型の分布となることからこの式が導かれる．全ての関係が無関連である場合は，小さい方の閾値が有効なので，各閾値が L 字型の分布となり，閾値 t_i が制約を与える関連度の数は，

$$2(N - i + 1) - 1 = 2N - 2i + 1 \tag{5}$$

となる．全ての関係が有関連であるとき，全関連度の中で閾値 t_i の制約を受ける関連度の率は，

$$(2i - 1) / N^2 \tag{6}$$

である．全ての関係が無関連であるとき，関連度の中で閾値 t_i の制約を受ける関連度の率は，

$$(2N - 2i + 1) / N^2 \tag{7}$$

である．分母が N^2 となっているのは，全ての語の組み合わせが N^2 となるためである．これらの式をグラフにすると図 2のようになる．関係が有関連のみであるときのグラフと無関連であるときのグラフは左右対称な直線となる（図 2）．有関連のグラフで i が 1 のとき $1/N^2$ となっている．これは，全ての関係が有関連のとき，全ての関連度の $100\times1/N^2$ %が，閾値 t_1 の制約により t_1 未満の値を取れないという意味である．t_1 未満でなければ良いので，実際には t_1 以上であれば何でも良い．無関連のグラフでは i が 1 のとき $(2N–1)/N^2$ となっている．これは，全ての関連度の $100\times(2N–1)/N^2$ %が，閾値 t_1 の制約により t_i 以上の値を取れないという意味である．

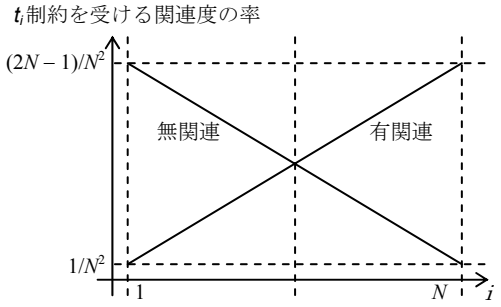

図2 閾値と制約を受ける関連度の率の関係

実際の状況において図 2のグラフがどうなるかについて考察する．結論から言えば，実際の状況における有関連のグラフと無関連のグラフはそれぞれ，図 2の有関連のみのグラフと無関連のみのグラフに一致する．全関係のうち有関連である率を p_a とすると，式(6)，式(7)からそれぞれ式(8)，式(9)ように導かれる．

$$\frac{p_a(2i+1)}{p_a N^2} = \frac{2i-1}{N^2} \tag{8}$$

$$\frac{(1-p_a)(2N-2i+1)}{(1-p_a)N^2} = \frac{2N-2i+1}{N^2} \tag{9}$$

分母と分子の両方に式(8)では有関連の率 p_a，式(9)では無関連である率 $1-p_a$ をかけている．式(8)，式(9)の分母はそれぞれ有関連の全関係数，無関連の全関係数であるため，全関係数 N^2 に p_a や $1-p_a$ を単純にかければ良い．

式(8)，式(9)の分子はそれぞれ閾値 t_i の制約を受ける有関連の関連度の数と無関連の関連度の数なので，分母のように単純にはいかない．しかし，分布が大きく偏る理由がないこと，全ての関係の数が極めて大きいことから式(8)，式(9)に問題はないと考える．分布が偏るには，関連度を計算する 2 語の閾値とそれらの語間の関係の強さに相関が必要であるが，基本的にそのような相関は存在しない．例えば，閾値が大きな語ほど有関連の語が増加したり，逆に減少したりすれば分布は偏るが，そのようになる理由は特に存在しない．また，分布に多少偏りがあっても，全ての関係の数は極めて大きいため，式(8)，式(9)の分母も分子も大きな値であり，全体としてはほとんど式(8)，式(9)に一致するはずである． 以上から，実際の状況における，閾値とそれが制約を与える関連度の率の分布は図 2と同じであることがわかる．

図 2から有関連と無関連をある程度識別できる閾値が存在し，最良の閾値が 2 直線の交点であることがわかる．N が奇数のとき閾値は中央の閾値 $t_{(N+1)/2}$ となる．有関連の関連度のうち $t_{(N+1)/2}$ 未満になってはいけない率は，i が 1 から $(N-1)/2$ の閾値 t_i に制約される関連度の率を 1 から引いた値となる．等差数列の合計の式を用いて次のようになる．

$$
\begin{aligned}
&1-\sum_{i=1}^{(N-1)/2}\frac{1}{N^2}(2i-1)\\
&=1-\frac{1}{N^2}\left\{1+2\left(\frac{N-1}{2}\right)-1\right\}\left(\frac{N-1}{2}\right)\frac{1}{2} \qquad (10)\\
&=\frac{3}{4}+\frac{1}{2N}-\frac{1}{4N^2}
\end{aligned}
$$

ただし，語数 N は非常に大きいため式(10)は 75%となる．無関連のうち $t_{(N+1)/2}$ 以上になってはいけない率は，グラフが対称なため，有関連の関連度のうち $t_{(N+1)/2}$ 未満になってはいけない率に一致する．したがって，閾値の数が奇数であるとき有関連と無関連を識別する閾値として閾値 $t_{(N+1)/2}$ を取れば，本稿で提案する識別率は 75%以上となる．ただし，ここでの識別率は 2 段階の関係の強さを扱っている．75%ではなく 75%以上となるのは，閾値 t_i の制約を受けた関連度は，有関連であれば t_i 以上となり，無関連であれば t_i 未満となるためである．例えば，もし，ある有関連の関連度が，i が$(N+1)/2$ 未満の閾値 t_i の制約を受けているとき，単に t_i 以上であるだけでなく閾値 $t_{(N+1)/2}$ より大きな値を取れば，識別率は 75%より大きくなる．無関連の場合であっても同様である．識別率の上限の制約は存在しないため 75%以上となる．また，語数 N が偶数であっても，以上と同様の結果となる．

75%と言う値はそれほど大きな値ではない．しかしながら，この値は，理論上の最低値であり，意図的に努力したとしても 75%の識別率で高精度な関連度を実現するのは極めて困難である．現実的には，80%や 90%の識別率が要求され，結局，全ての語の閾値を等しくなるのと変わらなくなる．

6 共通閾値の応用

共通閾値の存在は，関連度が関わる多くの場面に影響を与える．

6.1 関連度の利用

共通閾値の存在により，関連度を利用して語間の関係の強さを判断するときには，基本的に共通閾値との比較だけで判断ができることがわかる．また，十分に精度の高い関連度が計算できるという前提で，関連度を用いた語間の関係の強さを判断するシステムを構築するなら，語ごとに大きく閾値が異なることを考慮した設計は誤りである．

6.2 関連度の構築

共通閾値の存在は，関連度計算方式を構築する際の指針となる．基本的には，各語の関連度の閾値が計算の状況によって変化しないようにすれば良い．したがって，関連度の分布状況は，語ごとに変化しない方が良いと考える．

例えば，語の意味を語の集合で定義する知識ベースである概念ベースを用いた関連度計算[1]を例にすると次のようになる．この関連度計算方式は語の集合の一致の程度から関係の強さを定量化する方式であるが，関連度計算に用いる語数は理論上，関連度の分布に大きな影響を与える．このことと共通閾値の存在から，語ごとに異なる関連度計算に利用できる語の集合をそのまま全て使わず，例えば一定数にそろえた方が良いという予想ができる．

6.3 関連度の評価

識別率は，共通閾値の存在から導かれ，特定の関連度計算方式の特性や，特殊な仮定などに依存していないため，どのような関連度計算方式などにも適用できる汎用的な評価法である．

7 おわりに

本稿では，関連度が完全であるとき関連度の定義から論理的に，関連度を計算する語によって変わらない，関係の強さを識別できる閾値である共通閾値が存在していることを示した．また，共通閾値により強さを正確に識別された関係の率である識別率を，関連度の評価値として提案した．完全な関連度は関連度の目標であるため，その特性の把握は重要である．

共通閾値は，関連度の利用，関連度の構築，評価方法などの関連度が関わる多くの場面に応用可能であり，有用性が高い．本稿では観点のない単純な語間の関連度を用いて述べたが，共通閾値や識別率は適用範囲が極めて広く，文書などの語以外の関連度や観点付きの関連度など，関連度であれば多くのものに適用可能である．

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．

参考文献

[1] 渡部広一，河岡司：“常識的判断のための概念間の関連度評価モデル”，自然言語処理，Vol.8，No.2，pp.39-54，2001

[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