数式の規則的生成による類似尺度の探索の研究

皆川 歩
豊橋技術科学大学
知能・情報工学専攻

岡部 正幸
豊橋技術科学大学
情報メディア基盤センター

梅村 恭司
豊橋技術科学大学
情報・知能工学系

{minagawa@ss.cs, okabe@imc, umemura@ics}.tut.ac.jp

1　はじめに

本論文では，あるデータ集合に現れる事象の名前をラベルとし，ラベル同士の一対多関係を推定する問題を扱う．ここでの一対多関係とは，例えば新聞記事に現れる地名であれば，都道府県を表すラベルと市郡を表すラベルの関係などである．

文章中に現れる語句同士の関係を統計学的に分析することは，自然言語処理の標準的な技術である [1]．また，これまでに，データ集合に現れるラベル間の関係を推定する方法として，ラベルの出現パターンを用いる方法が提案されている [2, 3]．この方法では，ラベルの出現パターンの類似度を計算する尺度に何を用いるかが重要となるが，山本らの論文により，推定する関係が一対多関係であると先見的にわかっている場合，補完類似度を用いることが提案されている [4, 5]．

本研究では，類似尺度となる関数を，限定した範囲で規則的に生成し，精度の比較を行う．本稿では言語処理学会第17，18回年次大会の発表を元に，関数の探索範囲を拡大し先行研究での提案尺度との性能比較を行った [6, 7]．また既存の類似尺度よりも精度が良い関数が，探索範囲の拡大により発見でき，その比較評価が統計的に有意であることを示すものである．

2　問題定義

2.1　ラベルの一対多関係

本論文では，事柄を表す名前の総称をラベルと呼称し，ラベル間の関係を抽出する問題を取り扱う．

本論文における一対多関係とは，一階層の多分木で表現されるラベル要素の関係である．ここで仮に，一対多の「一」に対応するラベルを親ラベル，一対多の「多」に対応するラベルを子ラベルと呼ぶことにする．一対多関係が成り立つには2つの条件がある．最初の条件は，子ラベルは複数の親ラベルを持たず，必ず1つの親ラベルを持つことである．また，次の条件は，親ラベルは必ず複数の子ラベルを持つことである．図1に地名をラベルとした一対多関係の例を示す．

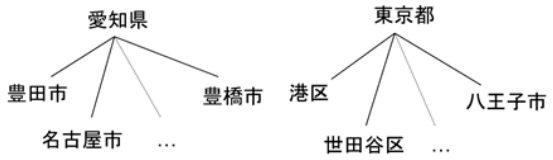

図 1: 一対多関係の例

2.2　ラベルの関係推定方法

これまでに，データ集合に現れるラベル間の関係を推定する方法として，ラベルの出現パターンの類似度を用いる方法が提案されている [2, 3]．ラベルの出現パターンをベン図によって表したものを図2に示す．

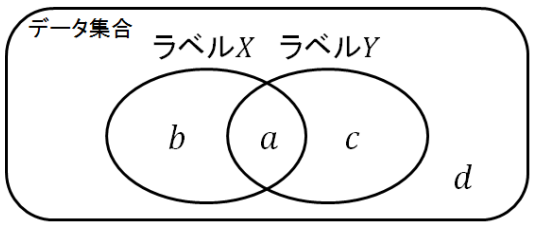

図 2: 出現パターンを表すパラメータ

- a：二つのラベルが同時に出現するデータ数
- b：ラベル X のみが出現するデータ数
- c：ラベル Y のみが出現するデータ数
- d：二つのラベルが出現しないデータ数

図 2 の 2 つの楕円はそれぞれのラベルが出現するデータ数を表す．a, d はラベルの組の一致度，b, c は不一致度を示す．

ラベルの関係を推定するには，まず，全てのラベルの集合から，2 つのラベルの組み合わせを取り出し，それぞれの組み合わせについてパラメータ a, b, c, d を求める．そして，もとめたパラメータを基に類似度のスコアを計算し，スコアの高いラベルの組ほど，関係性が強いと判断する．上記のパラメータから，ラベル間の関係性のスコアを求める関数が類似尺度である．

3 研究動向

3.1 提案手法

本稿では，次に示す 3 つの類似尺度に注目し，それらを含む範囲で類似尺度の探索を行う．

1 つ目は，補完類似度である．前節のパラメータを用いた定義を示す．

$$\text{補完類似度} = \frac{ad - bc}{\sqrt{(a+c)(b+d)}} \tag{1}$$

この類似尺度は，山本らの先行研究により，一対多関係の抽出に有効であることが示された [4]．

2 つ目は，ϕ 相関係数である．前節のパラメータを用いた定義を示す．

$$\phi\text{相関係数} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+c)(b+d)}} \tag{2}$$

この類似尺度は，統計における主要な相関係数として提案されている [1]．

3 つ目は，条件付き確率である．前節のパラメータを用いた定義を示す．

$$\text{条件付き確率} = \frac{a}{a+b} \tag{3}$$

また探索した関数の評価に際して，次に示す正規化補完類似度にも着目する．

$$\text{正規化補完類似度} = \frac{ad - bc}{\sqrt{(a+c+1)(b+d)(a+1)d}} \tag{4}$$

(1) と (2) を含む範囲で探索した関数の中では，(4) が高精度であることを [6] が示している．

関係抽出の対象ラベルには地名を選び，関数の探索と評価では，新聞記事データ 7 年分（毎日新聞 91〜97 年度版）から抽出した地名を，実データの集合として用いた．地名を選択した理由は，都道府県と市町村区群の地理的包含関係に一対多関係が成り立ち，また正解の組みが実世界で定まっているためである．

3.2 数式の生成

本研究では，類似尺度となる数式を生成し，一対多関係に有効な類似尺度を探索する．このとき，生成する式の候補は無限に存在するが，それら全てについて評価を行うことは不可能である．このため本論文では，生成する式の範囲を，3.1 で挙げた尺度の類型となる次の式の形式に限定する．

$$\frac{\alpha - \beta}{\sqrt{(\gamma)(\delta)(\epsilon)(\zeta)}}$$

分子部分の α は加算項，β は減算項であり，それぞれ出現パターンを表すパラメータ a, b, c, d から単数あるいは複数選択した項の積という形式である．a, b, c, d の指数は 1 のみに限定する．過去の探索では，分子部分は $ad - bc$ で固定であったが，条件付き確率を範囲に含むという条件の元，探索範囲の拡大を行った．

分母部分は平方根内に 4 つの括弧を持ち，それらの括弧が含む数式をそれぞれ変化させる．γ，δ，ϵ，ζ はそれぞれ，出現パターンを表すパラメータ a, b, c, d または定数 1 の中から，単数あるいは複数の項が選択され，選択された全ての項を加算する数式が入る．

上記の方針により，生成する数式の総数は 10,434,600 個となった．これらの数式には，3.1 で挙げた尺度を含む．

4 関数の探索とその評価

3.2 より，生成する数式の総数は 10,434,600 個となるが，それらすべてについて新聞記事データ 7 年分を用いた性能測定を行うことは現実的ではない．そのため，初めに小さなデータ集合で実験を行い，見込みの無い関数については計算対象から除外する．そして見込みのある関数についてのみ，次いで大きなデータ集合を用いて関数の性能測定を行う．

4.1 共起回数とスコアの関連測定

生成された関数について見込みがあるかどうかの判断基準として，共起回数に対するスコアの変動を用いた．これは，ラベルの共起回数が多いほど類似度を高く設定することが自然であると考えたためである．そのため，共起回数 (a) に対して類似度のスコアが単調

表 1: 代表に選んだラベル対とパラメータ

ラベル対	a	b	c	d
「大阪，大阪市」	1860	4473	1124	45872
「岩手県，盛岡市」	32	249	55	52993
「愛媛県，温泉郡」	13	202	0	53114

増加である関数を見込みが有る関数とし，代表的な正解のラベル対を用いてスコアを計算し，それらのラベル対のパラメータにおいて共起回数 (a) が増加した際にスコアが増加する関数についてのみ性能評価の対象とすることにした．

ここでは，代表的な正解のラベル対として「大阪，大阪市」,「岩手県，盛岡市」,「愛媛県，温泉郡」の 3 つを用いた.「大阪，大阪市」は，実データの正解ラベル対の中で最も共起回数が多いものである.「岩手県，盛岡市」は県名と県庁所在地という組み合わせの 1 つであり，実世界でも比較的関連が強いと考えられるラベル対である.「愛媛県，温泉郡」は完全な包含関係となるラベル対である．各ラベル対のパラメータを表 1 に示す.

上記の実験の結果，2,182,441 個の関数が候補として残り，以降の実験での計算対象とした.

4.2 1 年分の新聞記事データを用いた関数の探索

小さいデータ集合での探索で候補として残った関数に対して，91 年度版毎日新聞データを用いて精度の測定を行った.

この行程では，候補として残った関数を用いて，データ集合に現れるラベルの全ての組み合わせについて類似度の計算を行う．次いでそれらのラベルの組を，類似度のスコアが高い順にソートする．そして，この順位付けされたラベルの組から R-精度を計算し，各々の数式の性能指標とする.

実験結果を表 2 に示す．この表は 91 年度版毎日新聞データで計算を行い R 精度を比較した結果の上位 10 件である．R 精度比較で上位となった関数は，分子部分が $adb - bd$ であるもので占められていることが確認できる.

表 2: 1 年分のデータを用いた探索で高精度の関数

関数	R 精度
$\frac{abd-bd}{\sqrt{(a+1)(b+1)(a+b+1)(a+c+1)}}$	0.841
$\frac{abd-bd}{\sqrt{a(a+c)(b+1)(a+b+1)}}$	0.840
$\frac{abd-bd}{\sqrt{(a+b)(a+1)(b+1)(a+c+1)}}$	0.839
$\frac{abd-bd}{\sqrt{a(a+c+1)(b+c+1)(a+b+c+1)}}$	0.839
$\frac{abd-bd}{\sqrt{(a+1)(a+b+1)(a+c+1)(b+c+1)}}$	0.838
$\frac{abd-bd}{\sqrt{ab(a+c)(a+b+1)}}$	0.837
$\frac{abd-bd}{\sqrt{(a+1)(b+1)(a+c+1)(a+b+c+1)}}$	0.836
$\frac{abd-bd}{\sqrt{a(a+c)(a+b+1)(b+c+1)}}$	0.836
$\frac{abd-bd}{\sqrt{a(a+b+c)(a+c+1)(b+c+1)}}$	0.836
$\frac{abd-bd}{\sqrt{b(a+b)(a+1)(a+c+1)}}$	0.835

実験の結果，次の関数が最も良い精度を記録した.

$$\frac{abd - bd}{\sqrt{(a+1)(b+1)(a+b+1)(a+c+1)}} \quad (5)$$

また，精度が良く形式が特徴的な関数として次のものが発見できた.

$$\frac{abd - acd}{\sqrt{(a+1)(a+c+1)(b+c+1)(a+b+c+1)}} \quad (6)$$

上記の関数は，拡大した探索範囲に属する，新たなタイプの関数である.

4.3 6 年分の新聞記事データを用いた関数の評価

4.2 で得られた 2 つの関数と，補完類似度，正規化補完類似度について，関数探索に使用しなかった 6 年分の新聞記事データで性能比較を行った.

新聞記事データについては，それぞれの年度のデータを前半 (1 月から 6 月) と後半 (7 月から 12 月) に分割し，合計 12 個のデータで実験を行った．性能の指標には R 精度を用いた.

実験結果を表 3 に示す．下線のある数値はそれぞれの実験対象データでの最も良い精度の値である.

関数 5 は，6 年分の新聞記事データで比較すると，12 個中 4 個のデータで正規化補完類似度よりも良い精度を示し，2 個のデータで同精度，6 個のデータで正規化補完類似度より低い精度となっている.

表 3: 6 年分の新聞データを用いた R 精度の測定

データ	補完類似度	正規化補完類似度	関数 5	関数 6
92 年前半	0.683	0.690	0.690	0.700
92 年後半	0.454	0.687	0.679	0.670
93 年前半	0.513	0.746	0.737	0.753
93 年後半	0.386	0.683	0.683	0.686
94 年前半	0.458	0.693	0.666	0.734
94 年後半	0.295	0.670	0.663	0.676
95 年前半	0.221	0.597	0.431	0.621
95 年後半	0.326	0.661	0.651	0.707
96 年前半	0.455	0.677	0.701	0.697
96 年後半	0.506	0.612	0.659	0.658
97 年前半	0.430	0.650	0.684	0.662
97 年後半	0.414	0.562	0.650	0.597

また関数 6 が 12 個中 11 個のデータで正規化補完類似度よりも良い精度を示していることが確認できる.

5 考察

5.1 評価の有意差

関数 6 と正規化補完類似度の比較結果に有意性が認められるか，符号検定を行う.

次の統計的仮説をたてる.

H_0 : 関数 6 と正規化補完類似度の精度に有意差が有る.

H_1 : 関数 6 と正規化補完類似度の精度に有意差が無い.

このとき，R-精度の分布が等しいという仮定の下で，優位確率は次式のようになる.

$$P = \frac{2({}_{12}C_0 + {}_{12}C_1)}{2^{12}} = 0.006348\ldots < 0.01$$

これにより，補完類似度，正規化補完類似度と関数 6 の比較結果は，危険率 1 % で統計学的に有意であると言える.

6 まとめ

本研究では，ラベルの関係抽出問題に対して，補完類似度と ϕ 相関係数，条件付き確率の類型を範囲として類似尺度の関数を規則的に生成し，性能測定を行った．これにより，1 年分の新聞記事データでトップの精度を示す関数 5 と精度が良く形式が特徴的な関数 6 を発見した.

またそれらの関数と補完類似度，正規化補完類似度の間で性能比較を行い，探索に用いたデータとは異なる 6 年分の新聞記事データで関数 6 が既存の類似尺度よりも良い性能を示すことを確認した．関数 6 の数式を以下に示す.

$$\frac{abd - acd}{\sqrt{(a+1)(a+c+1)(b+c+1)(a+b+c+1)}}$$

また，関数 6 と正規化補完類似度の比較結果について符号検定を行い，関数 6 の方が有意に精度が優れていることを示した.

参考文献

[1] Christopher Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[2] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pp. 32–41, New York, NY, USA, 2002. ACM.

[3] S. Choi, S. Cha, and C. C. Tappert. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, Vol. 8, No. 1, pp. 43–48, 2010.

[4] 山本英子, 梅村恭司. コーパス中の一対多関係を推定する問題における類似尺度. 自然言語処理, Vol. 9, No. 2, pp. 45–75, 2002.

[5] 澤木美奈子, 萩田紀博. 補完類似度に基づく新聞見出し文字の領域抽出と認識. 電子情報通信学会技術研究報告. PRU, パターン認識・理解, Vol. 95, No. 278, pp. 19–24, 1995-09-28.

[6] 皆川歩, 岡部正幸, 梅村恭司. 数式の網羅的な生成による新たな類似尺度の発見と評価. 言語処理学会第 17 回年次大会発表論文集, 2011.

[7] 皆川歩, 岡部正幸, 梅村恭司. 数式の網羅的な生成による新たな類似尺度の決定とその評価. 言語処理学会第 18 回年次大会発表論文集, 2012.