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

皆川 歩
豊橋技術科学大学
情報工学課程

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，ラベル同士の関係を推定する問題を扱う．このようなラベルの関係抽出の問題では，ラベル同士の関係は一対多関係になることが多いと知られている．ここで，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について，相関係数などの関数を用いて統計学的に分析することは，Manningらがその著書において，自然言語処理の標準的な技術として説明している [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の類似度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どのようなものを用いるかが重要となるが，山本らの論文により，先見的に推定する関係が一対多関係であるとわかっている場合に，ラベル間の関係推定に補完類似度を用いることが提案されている [4, 5]．

実験対象となる事象としては，地名 (都道府県市郡名) を用いた．これは，地名は実世界の地理関係により一対多関係をもち，また正解となる組み合わせも実世界で定まっているためである．実験の対称とするデータ集合には，実際の新聞記事を用い，地名の出現パターンから一対多関係を推定する能力を測定した．

本論文では，類似度を判断するための尺度となる数式を，限定した範囲で網羅的に生成し，既存の類似尺度と精度の比較を行う．その結果，既存の類似尺度よりも良い精度を示す尺度を発見し，これを本論文における提案尺度とすることを報告する．また，提案尺度と補完類似度を比較し，提案尺度が既存の類似尺度よりも良い精度を示した理由について考察を行う．

2　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．

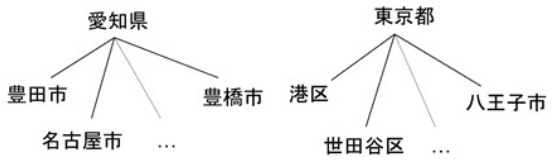

図 1: 一対多関係の例

3　ラベルの関係の推定方法

ラベル間の関係を推定するには，ラベルの出現パターンを利用する．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．

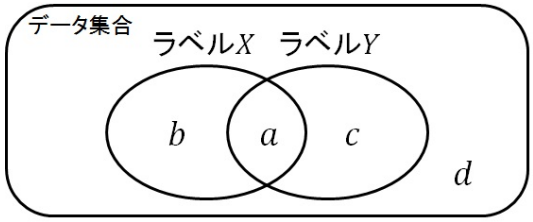

図 2: 出現パターンを表すパラメータ

- 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り，この類似尺度によって，関係推定の性能は大きく変化する．

4 主要な類似尺度

本論文では，過去の研究において提案された 2 つの類似尺度に注目する．

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は次のようになる．

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

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は次のようになる．

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

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].

これらの 2 つの式は，分子部分は共通しており，一致度 ad が大きく，不一致度 bc が小さいほど類似度のスコアは増加する．2 式で異なるのは分母の数式である．$\sqrt{(a + c)(b + d)}$ はどちらの式にも共通しているが，ϕ 相関係数には共通部の他に $\sqrt{(a + d)(b + c)}$ が存在する．

2 つのパターンの類似度を求める場合，パターンの一致している部分のみに着目する類似尺度を用いることが多いが，これらの尺度はパターンを入れ替えても同じ値を持つ．このような性質を持つ類似尺度は対称性を持つといい，ϕ 相関係数は対称性を持つ．これに対し，補完類似度は分子のパラメータ b,c により，2 つのパターンの相違を考慮した定義である．このため補完類似度は非対称性を持つといえる．山本らの先行研究により，非対称性を持つ補完類似度の方が，他の対称性を持つ類似度よりも一対多関係推定に有効であることが示されている．

5 数式の生成

本論文では，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，補完類似度と ϕ 相関係数の類型となる次の式の形式に限定する．

$$\frac{ad - bc}{(?)(?)(?)(?)}$$

数式の分子部分は前節で示した類似尺度と同一である．一致度 ad が大きく，不一致度 bc が小さいほど類似度のスコアは増加する．

数式の分母部分は，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．平方根内のそれぞれの括弧は，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．このとき，a, b, c, d の係数は 1 のみに限定する．これは，係数に 1 以外も含むと，数式のパターンが増大し，評価が現実的には困難になるためである．このため平方根内の数式は項の有無のみが変化する．

上記の方針により，生成する数式の総数は 46376 個となった．これらの数式には，補完類似度と ϕ 相関係数を含む．

6 評価実験

6.1 対象データ

本論文では，関係抽出の対象ラベルとして日本の地名を選んだ．地名を選択した 1 つ目の理由は，地名間には実際に一対多関係が成り立つためである．例えば，県名と市名には地理的な包含関係があり，これは一対多関係である．2 つ目の理由は，正解となる地名の組み合わせが実世界において定まっているためである．これらの理由から，日本の地名をラベルとし，推定する一対多関係は地名間の地理的な包含関係とした．

表 1: 提案尺度と既存の類似尺度の精度比較

年度	91	92	93	94	95	96	97	平均
提案尺度	0.747	0.807	0.833	0.793	0.800	0.793	0.749	0.789
補完類似度	0.699	0.559	0.431	0.362	0.315	0.485	0.416	0.467
相互情報量	0.529	0.371	0.287	0.256	0.190	0.322	0.285	0.320
信頼度	0.024	0.123	0.158	0.061	0.239	0.220	0.232	0.151
Φ相関係数	0.001	0.055	0.103	0.039	0.113	0.128	0.146	0.084

実験対象のデータ集合には，新聞記事データ 7 年分（毎日新聞 91〜97 年度版）を用い，都道府県市郡の地名を抽出した．また，正解となるラベルの組み合わせはポスタルガイドから抽出した．これにより，ラベルが取りうる全ての組み合わせは 8272278 組となり，正解組数は 1239 組となった．

データ集合に現れるラベルの中には「大阪」などのように都道府県市郡を表す語がついていないものも含まれる．このような場合には,「大阪府」,「大阪市」というように都道府県市郡を表す語を補い，それにより正解の組み合わせになるのであれば，正解とする．

6.2 性能評価の方法

実験では，生成された各々の数式を用いて，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．

R-精度とは，関係推定の対象としたラベルの全正解組数を R とし，類似度のスコアにより順位付けされたラベルの組から上位 R 組を取り出し，その中に含まれる正解の割合が R-精度である．次に R-精度の定義を示す．

$$R = \text{全正解数}$$

$$\text{R-精度} = \frac{\text{上位 } R \text{ 組の中で正解関係にある組数}}{R}$$

6.3 実験結果

次の数式が 7 年分の新聞記事データ全てで安定して高い精度を示した．

$$\frac{ad - bc}{(a + c + 1)(b + d)(a + 1)d}$$

このため，上記の式を本研究における提案尺度とする．また，毎日新聞記事データ 91 年から 97 年度の各版における，提案尺度と既存の類似尺度との R-精度の比較結果を表 1 に示す．7 年分全てのデータで，提案尺度が既存の類似尺度よりも精度が良いことが確認できる．

提案尺度は補完類似度よりも，R-精度が平均で 0.322 だけ良い．また，補完類似度は各版の R-精度の最高値と最低値の差が 0.384 であることに対し，提案尺度ではその差は 0.086 と非常に小さく精度が安定している．

7 考察

7.1 実験結果の有意差

6 節の実験結果において，提案尺度は 7 年分の新聞記事データ全てで既存の関数よりも良い R-精度の値を示した．この結果に有意差が認められるか，符号検定を行う．

次の統計的仮説をたてる．

H_0：提案尺度と既存の関数の精度に有意差は無い．

H_1：提案尺度と既存の関数の精度に有意差が有る．

このとき，R-精度の分布が等しいという仮定の下で，7 個のデータセット全てにおいて，既存の関数の R-精度が提案尺度の R-精度を上回らない確率は次のようになる．

$$\frac{1}{2^7}({}_7C_0) = 0.007812\ldots < 0.01$$

これにより，前節の実験結果は，危険率 1 % で統計学的に有意であると言える．

7.2 式と補完類似度の比較

補完類似度の数式は，分子の次元数が分母の次元数よりも大きい．このため，データ数が多くなるとラベ

表 2: ラベルの組と出現パターンの例

ラベル X	ラベル Y	a	b	c	d
兵庫県	西宮市	1500	8042	190	67914
千葉県	松戸市	117	1175	6	76348
大阪	寝屋川市	145	26011	23	51467
大阪府	寝屋川市	154	4238	14	73240
大阪府	泉南	43	4349	12	73242

ルの出現頻度が等しくても，類似度のスコアは増加する．これは，補完類似度はデータ数が多くなるほど分子の式 $ad - bc$ の値を重視する傾向にあるとも言える．

また，ラベルの出現パターンを表すパラメータは，その傾向として，d が他のパラメータより大きな割合を示すことが多い．表 2 にラベルの組と出現頻度のパラメータの例を示す．このため，データ数が多くなるとラベルの出現頻度が等しくても ad は bc と比較し非常に大きな値となる．

この 2 つの事柄から，補完類似度はデータ数が増加すると ad を重視するようにバイアスがかかると考えられる．

ここで，提案尺度と補完類似度の数式を比較すると，提案尺度は補完類似度の分母に $\sqrt{(a+1)d}$ を追加した形に類似している．提案尺度はこの追加部分によりデータ数の増加によるバイアスを防止し，相対的に補完類似度よりも b または c の小ささを重視し，一対多関係のラベルが持つ包含関係に適した形になったと考えられる．

本論文の提案尺度は，補完類似度に対してデータ数に関する正規化を行った関数とも言える．このため，本論文における提案尺度を正規化補完類似度と命名する．

8 まとめ

本研究では，類似尺度の数式を限定した範囲内で網羅的に生成し，既存の類似尺度との性能比較を行った．これにより，既存の類似尺度よりも良い精度を示す関数を発見し，これを正規化補完類似度として提案した．正規化補完類似度は，既存の類似尺度よりも良い性能を示し，これは危険率 1 % で統計的に有意である．

9 今後の課題

本論文では，関係推定の対象データとして毎日新聞記事データ 7 年分のみを用いた．しかし，本論文で提案した正規化補完類似度の有効性を検証するためには，毎日新聞以外の新聞記事データ，あるいは地名以外の事象をラベルとした全く異なるデータを用いて評価実験を行う必要があると考えられる．

また，正規化補完類似度が一対一関係に対しどのような性能を示すかについては現段階では不明である．そのため，一対一関係推定の問題に対しても正規化補完類似度が利用出来るのか，検証を行う必要がある．

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.