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

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関係を推定する問題を扱う．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．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 [2]．

本研究では，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った．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，これを正規化補完類似度と命名し提案尺度とした．

本稿は言語処理学会第17回年次大会の発表を元に人工データを用いた実験と正規化補完類似度のスムージング項についての実験を行い，関数の生成について考察を行うものである．

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．

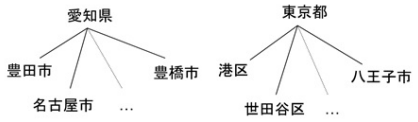

図 1: 一対多関係の例

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．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．

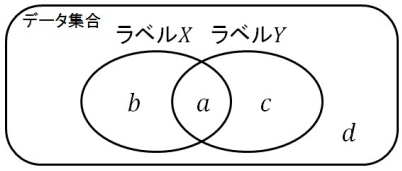

図 2: 出現パターンを表すパラメータ

- 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

本研究では，過去の研究において提案された 2 つの類似尺度に注目した．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．

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

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 [2]．

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．

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

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]．

本研究では，補完類似度と ϕ 相関係数の類型となる数式を網羅的に生成し，一対多関係に有効な類似尺度を探索した．

また，関係抽出の対象ラベルとしては日本の地名を選び，性能評価では，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，地名間には都道府県と市町村区群の地理的包含関係があるため一対多関係が成り立ち，また正解の組み合わせが実世界で定まっているためである．

3.2 正規化補完類似度

次の数式が生成した関数の中では最も良い精度を示し，関数の生成範囲内には無い既存の類似尺度との比較でも最も良い精度であった．

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

よって，この式を本研究における提案尺度とし，正規化補完類似度と命名した．

3.3 補完類似度との比較

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ため，分母式よりも分子式 $ad - bc$ の値を重視する傾向にある．

また，ラベルの出現パターンを表すパラメータは，d が他のパラメータより非常に大きな値を示す傾向にあり，このため ad は bc と比較し非常に大きな値となりやすい．

この 2 つの事柄から，補完類似度は 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，正規化補完類似度と補完類似度の数式を比較すると，正規化補完類似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．正規化補完類似度はこの追加部分によりバイアスを防止し，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．

4 評価実験

4.1 人工データを用いた実験

正規化補完類似度が良い性能を示した理由についてさらに検討するため，この節では，正規化補完類似度が良い精度を示すデータモデルを予想し，そのモデルに従うデータを人工的に生成し，正規化補完類似度と補完類似度の振る舞いの違いを観察する．

ここで用いたデータのモデルは，ラベルの出現頻度に偏りがあり，雑音を含むデータモデルである．

ラベルの出現頻度に偏りを持たせる理由は，正規化補完類似度の方が，互いに出現頻度が高いが正解関係では無いラベル対を正しく判別する能力に優れていると予想できるためである．

出現頻度が高いラベル同士は正解関係でなくとも a,b,c のパラメータが大きくなりやすく，補完類似度は a の値を重視するためこれを正解関係と誤認しやすい．対して正規化補完類似度は c の小ささをより重視するため，これを正解関係ではないと正しく判別できると考えられる．

出現頻度に偏りがあるデータにおいて正規化補完類似度が補完類似度よりも良い精度を示すならば，正規化補完類似度の性質についてのこれらの予測が正しく，

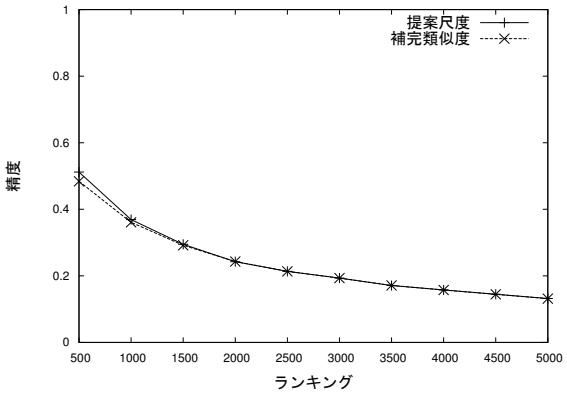

図 3: 人工データでの精度

またそれが実データでも良い精度を示した理由であると考えられる.

このモデルに従うデータは，例えば { 静岡県，清水市，神奈川県，横浜市，京都府 } というデータである. これは，正解である関係集合から〈静岡県，清水市〉と〈神奈川県，横浜市〉の２つを取り出し，これに雑音「京都府」を追加した集合である.「京都府」はデータ内の他のラベルとは正解集合に含まれる関係を持たない.

正解ペアは高頻度ペア 15 組と低頻度ペア約 1000 組に分けた．そして，高頻度ペアと低頻度ペアの出現が等確率になるようにした.

このように作成したデータを 1000 個持つデータ集合を作成した.

実験結果を図 3 に示す．このグラフは，上位の組の数を横軸，そのときの精度を縦軸とする．上位に位置する組において正規化補完類似度が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.

また，同様の方法でデータを 7 個作成し実験を行ったが，それらすべてで，正規化補完類似度の方がランキング上位の組での精度が良いという結果が得られた.

4.2 スムージング項による影響の測定

正規化補完類似度と補完類似度の分母を比較すると，正規化補完類似度では変数 a を含む項に定数 1 が加えられていることがわかる．この定数 1 は変数 a の値が 0 であるとき類似度の計算が不可能になることを防ぐためのスムージング項であるといえる.

本研究では，関数の探索範囲の都合から，項に含まれる定数は 1 に限定したが，これは評価する数式の総数を現実的に計算可能な範囲に限定するための制限である．スムージングと解釈される項があるのでこれをスムージング項であると考えると，定数 1 がスムージングに適した値かどうかは不明であり，他に最適な値が存在する可能性も考えられる.

このため，本節では正規化補完類似度のスムージング項を変化させ性能測定を行った．対象データには新聞記事データを用い R 精度を性能指標とした．実験結果を表 1 に示す.

実験結果より，対象データによって精度がピークとなるスムージング項の値は異なり，定数 1 は各データで精度のピークとなったスムージング項の値の範囲内 (0.8〜5) にあることが確認できた．また，スムージング項の値はそれほど敏感なパラメータではないと言える.

スムージング項をどのような値に設定すれば良いかは今後の検討課題であるが，データにより精度が最も良くなるスムージング項の値が異なるため，現状では最適な値を決定できるかは不明である．しかし定数 1 に設定した場合，多くの年度のデータでピークに近い精度が測定されており，さほど悪い結果では無いといえる.

5 考察

関数を多量に生成することにより，本研究で実験に用いた実データに対してのみ良い性能を示す関数が偶然得られたのではないかという議論もある．つまり，実験において生成した関数が多岐に渡り，それらの一部が偶然良い性能を示した可能性があるという議論である.

しかし，評価実験の成績でトップに類する関数は全て正規化補完類似度の類型で占められ共通の性質を持っており，それらの共通の性質を持つ関数群についても解釈と説明は可能である.

また，人工データにおいて補完類似度よりも精度が良いことは，補完類似度の問題を正しく改良していることと解釈できる.

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 相関係数の類型のみに限定した範囲内で類似尺度の数式を網羅的に生成し，性能測定を行った．これにより，7 年分の全ての新聞記事データでトップかあるいはそれに次ぐ精度を示す関数を発見し，その関数を正規化補完類似度として提案した．正規化補完

表 1: スムージング項による精度の変化

スムージング項	年度						
	91	92	93	94	95	96	97
0.6	0.739	0.770	0.755	0.712	0.778	0.739	0.723
0.7	0.743	0.801	0.783	0.764	0.785	0.743	0.744
0.8	0.747	0.801	0.831	0.790	**0.805**	0.756	0.745
0.9	0.746	0.803	0.832	0.791	0.804	0.758	0.747
1	0.747	0.807	0.833	0.793	0.800	0.793	0.749
2	0.786	0.845	**0.835**	**0.806**	0.737	**0.821**	0.789
3	0.801	0.839	0.811	0.794	0.654	0.812	**0.810**
4	0.807	**0.850**	0.810	0.776	0.609	0.797	0.790
5	**0.833**	0.842	0.793	0.744	0.575	0.781	0.772
6	0.826	0.840	0.784	0.727	0.549	0.764	0.749
7	0.832	0.820	0.752	0.701	0.520	0.751	0.736

類似度の数式を以下に示す.

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

また正規化補完類似度と既存の類似尺度の性能比較を行い，7 年分全てのデータで正規化補完類似度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.

正規化補完類似度が既存の類似尺度よりも良い性能を示した理由について，元となった数式の 1 つである補完類似度との比較を交え考察し，山本らの実験において高い精度を示した補完類似度に対しバイアスを取り除く正規化を行っているためであると解釈した.

また，実データでの実験結果と 3.3 節での考察から，正規化補完類似度が良い精度を示すと予想されるラベルの出現頻度に偏りがありノイズを含むモデルに従うデータを人工的に生成し，ランキング上位のラベル対で正規化補完類似度が補完類似度よりも良い精度を示すことを確認した.

さらに，正規化補完類似度のスムージング項について実験を行い，スムージング項の値によって性能が変化することを確認したが，最適な値が決定できるかについては現状では不明である.

正規化補完類似度が，多量に生成された関数のうち，偶然本研究で行った実験に対してのみ良い精度を示したものではないかという議論があるが，これに対し，補完類似度との比較及び人工データでの実験を通して正規化補完類似度が問題に対して正しく改良されている理由を示した.

7 今後の課題

本研究では，関係推定の対象データとして，実データでは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.

分布類似度における文脈と単語の関係を一対多関係とみなしたとき，それらに正規化補完類似度の適用が可能かどうか調査し，その評価を行うことも今後の課題として挙げられる.

参考文献

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

[2] 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.