補完類似度による地名情報の抽出

1T－9

山本英子　松本兼一　梅村恭司
豊橋技術科学大学　情報工学

1　はじめに

現在，様々な分野で情報の電子化が進み，巨大なデータが存在するようになっている．また，ネットワークの技術の進歩や記憶装置の低価格化により，誰もが巨大なデータに触れることのできる環境になりつつある．しかし，巨大なデータのすべてが有用な情報であるとは限らない．そこで，巨大なデータから有用な情報や隠れた情報を取り出す技術，データマイニングが重要となってきている．本研究では，実際にデータマイニングを含めたKDD(Knowledge Discovery in Databases)処理をデータから行なうことを試みた．本稿では，補完類似度による情報の抽出方法を提案する．

2　新聞記事と地名

本研究では，新聞記事をデータとして扱うことにした．新聞記事には，あらゆる品詞の語が出現するが，ある範囲の語に関しての情報に注目すれば良いと考えた．そこで，注目する語を選定するが，固有名詞というのは，特定のものの名称を表しているので注目する価値があると考えられるため，本研究では固有名詞に着目した．固有名詞の中の地名は，ある出来事が起こった場所として新聞記事に必ず出現するものであり，正解が判定できるので注目した．

3　地名の階層関係

これまでに共起情報を元にした情報の抽出やその応用は多く行なわれているが，出現頻度の情報の包含に視点を置き情報を抽出する研究はあまり行なわれていない．そこで，地名の出現パターンの包含状態より得られる地名の階層関係を取得することを考えた．本研究では，出現パターンより情報を抽出することに焦点を当て，階層関係を出現パターンの比較のみで取得することを試みた．本研究では，階層関係を補完類似度と相互情報量を用いて取得した．

4　補完類似度

補完類似度 [1] とは，パターン認識に用いられる関数のことである．これは，0,1 の 2 値で表される二つのベクトルの類似度を得ることができる．また，補完類似度は非対称性を持つ．入力は，2 値 n 次元のベクトル$\vec{F}=(f_1,f_2,...,f_i,...,f_n)(f_i=0$ または $1)$ と$\vec{T}=(t_1,t_2,...,t_i,...,t_n)(t_i=0$ または $1)$ である．$\vec{F}$の$\vec{T}$に対する補完類似度 S_cは次のように表される．

$$S_c(\vec{F},\vec{T})=\frac{a\cdot d-b\cdot c}{\sqrt{T\cdot(n-T)}}$$

$$a=\sum_{i=1}^{n}f_i\cdot t_i,\quad b=\sum_{i=1}^{n}f_i\cdot(1-t_i),\quad c=\sum_{i=1}^{n}(1-f_i)\cdot t_i,\quad d=\sum_{i=1}^{n}(1-f_i)\cdot(1-t_i),$$

$$T=\sum_{i=1}^{n}t_i,\quad n=a+b+c+d$$

S_cは$-\sqrt{T(n-T)}\leq S_c\leq\sqrt{T(n-T)}$の値域をとる．本研究では，各地名に対して，その地名が各記事に出現するかしないかを表す 2 値パターン列を作り，これをベクトルとして類似度を計算した．

Extracting Hierarchical Pairs of the Place Names using Complementary Similarity Measure
Eiko Yamamoto, Kenichi Matsumoto, and Kyoji Umemura
Department of Information and Computer Sciences, Toyohashi University of Technology

5 相互情報量

相互情報量は，二つの確率変数の依存性の度合を表す尺度である．これは，二つの確率変数の依存の度合が高くなるにつれて大きくなり，互いが独立な時，最小値0をとる．この尺度は一般的に効果が高い．

6 実験方法

実験は標準的なKDDプロセスに沿って行なった．

1. データの選択：新聞記事を形態素解析システム「茶筌」を用いて解析する．システムには日本の地名辞書を作成し加えておいた．形態素解析の結果から，地名と分類された固有名詞を日付と記事番号とともに抽出する．
2. データのコード化：それぞれの地名に対して，出現する記事には1，出現しない記事には0を割り当て，パターン列を作る．例えば，五つの記事のうち，ある地名が記事番号1,3,4に出現する場合，ある地名に対するパターン列は“10110”となる．
3. データの補強：それぞれの地名について，出現回数を数える．
4. データマイニング
   出現する全ての地名の組合せに関して，補完類似度または相互情報量を計算し，値が閾値より高いものを有用な情報とした．

7 閾値の決定

補完類似度を用いて地名の間の階層関係を取得する場合，全ての組合せを考慮する必要がある．その結果，補完類似度の高いものから出現パターンの類似性が強く，かつ階層関係が表れる地名の組が得られる．しかし，得られる全ての組合せが正しいわけではない．そこで，ある閾値を実験的に求め，それ以上の値をもつものをシステムの出力とした．実データを調査した結果，閾値を求める関数は次の式と推定できた．

$$\text{閾値}(n) = \frac{0.4}{2^{(\log_{10} n)}}$$

この関数を用いて，本実験での閾値を決定した．実験データとなった新聞記事数は53788であるから，閾値は0.01506となる．本研究では閾値より大きな補完類似度をもつ階層関係を有効と考えた．

8 実験結果の比較

これらの手法により取得した階層関係を再現率と適合率によって，評価し比較した．再現率と適合率は次の式で得られる．

$$\text{再現率} = \frac{\text{取得した正しい階層関係の数}}{\text{全ての正しい階層関係の数}},\quad \text{適合率} = \frac{\text{取得した正しい階層関係の数}}{\text{取得した全ての階層関係の数}}$$

実際の数値を代入すると，全ての正しい階層関係の数は1239，取得した全ての階層関係の数は642，取得した正しい階層関係の数は490であったので，再現率，適合率はそれぞれ次のようになる．

$$\text{再現率} = \frac{490}{1239} = 0.395,\quad \text{適合率} = \frac{490}{642} = 0.763$$

これを基準として他の方法と比較する．相互情報量を用いた手法の再現率と適合率は，上位642組において検査した．取得した全ての階層関係は642，取得した正しい階層関係の数は358であったので，再現率，適合率はそれぞれ次のようになる．

$$\text{再現率} = \frac{479}{1427} = 0.336,\quad \text{適合率} = \frac{479}{593} = 0.808$$

以上の結果から，補完類似度を用いた手法の方が高い適合率を示した．

[1] 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 信学技報 PRU95-106, pp.19–24, (1995).