ブログテキストにおけるトレンドパターンの学習と予測

学籍番号：90156039　谷内田研究室　小野 友也

1　はじめに

近年のインターネットの普及に伴い, ブログ記事を書く人も増加し，ブログ記事から有益な情報を抽出しようという試みが盛んに行われている．そこで本研究では, 単語の頻度の時系列変化の波形で定義されるトレンドパターンを学習し，トレンドを予測する手法を提案する．

2　学習方法

ブログ記事のテキスト部分に対し，形態素解析を行い，名詞，未知語を抽出する．次に，抽出された単語の頻度の時系列変化の情報，即ち，トレンドパターンをデータベースに登録する．データベースに登録されたトレンドパターンのうち，平均頻度よりも頻度の高い特徴的なトレンドパターンを検出し，多項式近似を行い，クラスタリングをすることで，トレンドパターンを学習する．

2.1　特徴的なトレンドパターンの検出

特徴的なトレンドパターンの検出には，χ^2 検定を用いる．トレンドパターンを検出しようとする着目区間の始まりの日を i, 区間幅を n, 単語の頻度を tf_k, 着目区間外の平均を μ としたとき，χ^2 統計量は以下の式で定義される．

$$chisq(i) = \sum_{k=i}^{n+i} \frac{(tf_k - \mu)^2}{\mu} \tag{1}$$

信頼係数 90% をみたす χ^2値を閾値 th として，$chisq(i) > th$ が成り立つ区間に特徴的なトレンドパターンが含まれているとする．この着目区間を時間軸にそってスライドさせる事で特徴的なトレンドパターンを検出する．

2.2　トレンドパターンの多項式近似

頻度にはブログ記事の収集量が一定でないためにノイズが含まれており，頻度をそのまま学習に用いると過学習の恐れがある．そこで，AIC を基準にした Gauss-Newton アルゴリズムを用いた多項式近似によりトレンドパターンを近似し，近似された頻度を要素とするパターンベクトルを生成する．

2.3　トレンドパターンの分類

階層的クラスタリングによってトレンドパターンを分類する．パターンベクトル $\mathbf{x}$ からみた パターンベクトル $\mathbf{y}$ との距離 $d_{\mathbf{x}\to\mathbf{y}}$, および, パターンベクトル $\mathbf{x}, \mathbf{y}$ 間の距離 $d(\mathbf{x}, \mathbf{y})$ を χ^2 統計量を用いて次式で定義する．

$$d_{\mathbf{x}\to\mathbf{y}} = \sum_{k=1}^{n} \frac{(y_i - x_i)^2}{x_i} \tag{2}$$

$$d(\mathbf{x}, \mathbf{y}) = (d_{\mathbf{x}\to\mathbf{y}} + d_{\mathbf{y}\to\mathbf{x}})/2 \tag{3}$$

式 (3) の距離が近い二つのクラスタをまとめていくことで，クラスタリングを行う．

3　予測方法

入力としてトレンドパターンが与えられたとき，学習の際と同様に多項式近似する．この入力は図 1 の比較区間 (comparision period) に相当する．学習したトレンドパターンと比較し，比較区間において式 (3) で定義される距離が小さい順に 5 つのトレンドパターンを出力する．

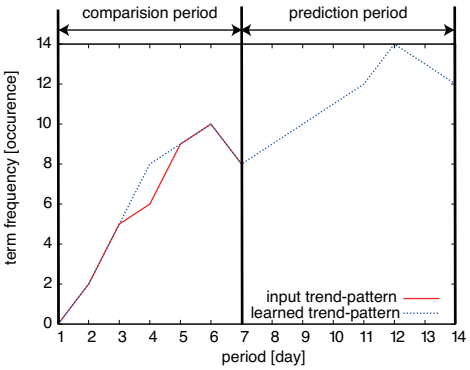

図 1　トレンドパターンの予測方法

4　実験

61,962 個のトレンドパターンを学習に用い，5,905 個の特徴的なトレンドパターンを学習した．比較区間の長さを p[日] とし，$p = 7, \cdots, 13$ と変化させ，約 1100 個の入力を与えた．比較手法として株価の予測に用いられる ARMA モデルを用いた．予測区間の平均二乗誤差を基準として，提案手法による予測のうち最も予測誤差の小さいものと ARMA モデルでの予測誤差のどちらが勝っているかをカウントした．結果を図 2 に示す．提案手法は予測日数が 2 日までと短い場合には，ARMA モデルよりも予測精度が高いことが分かった．

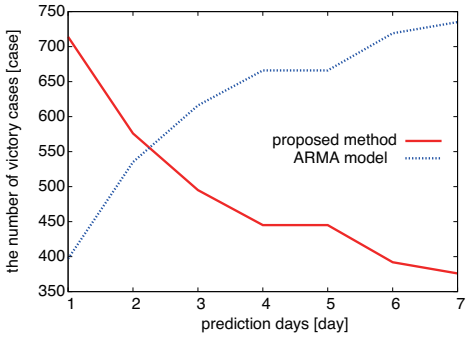

図 2　入力日数を変化させた時の，2 つの提案手法での誤差が小さかった単語数

5　まとめ

本研究では，トレンドパターンを学習し，予測を行う手法を提案した．ARMA モデルと比較を行ったところ，提案手法では予測日数が 2 日までならば ARMA モデルよりも予測の精度が高いことが分かった．今後はより長い予測日数でも精度よく予測を行うため，単語の共起関係を取り入れた予測方法へと改善していく．