時系列テキストデータを用いた多重スケールでのトピックモデルによる文書分類

1X10C044-1　郡司　巧
指導教員　後藤 正幸

1　研究背景・目的

近年，情報技術の発達に伴い，大量の文書データが日々生成されている．これらのデータは人手によるカテゴリ分類が困難であり, 自動文書分類の技術が必要とされている．また文書データそのものの多様性も増加し，Twitter や新聞記事のように連続的に蓄積される時系列データが存在する．このような時系列データを分析することで，日々の話題の変化やトレンドの移り変わりを抽出することは，例えば時系列データにおいて短期間でのトレンドを抽出することで文書分類への応用に役立つという可能性がある．

話題など様々なデータに隠れた潜在的なトピックを推定するモデルとしてトピックモデルがある．その中でもベイズ統計を用いて，時系列データを逐次的に分析し，潜在トピックを抽出する手法として MDTM の有効性が示されている．MDTM はトピック毎の単語分布を，複数の時間スケールを考慮した事前分布を仮定することで生成する．時間スケールとは，トピックが持つ単語分布の時間単位を指す．しかし，従来の MDTM では，全ての時間スケールでの単語分布を平滑化し，トピックでの単語分布としている．そのため時系列変化によってトピックが変化する単語を含む文書に対し，適切なトピックの単語分布を出力できないという問題点がある．さらに，MDTM を用いて文書分類を行い，カテゴリを抽出することを考える．しかし，MDTM はトピック毎の単語分布を出力する生成モデルであり，その単語分布はカテゴリ情報を持たないため，文書分類に適用することは困難である．

そこで本研究では，前者の問題に対しては，モデルの事前分布に対して，スケールの長さと単語出現傾向の違いによるパラメータ調整方法について提案する．また，後者の問題に対しては，MDTM を文書分類へ適用する手法を提案する．実際の新聞記事データを用い，過去の文書で学習を行い，新規入力文書データを現在の文書として，文書分類を行うことで提案手法の有効性を示す．

2　Multiscale Dynamic Topic Model

2.1　モデル

MDTM は逐次的に増加する文書集合の時系列変化を，複数の時間スケールを用いることで，考慮したトピックモデルである．時刻 t のある文書集合を $\boldsymbol{D}_t = \{d_{t,1}, \cdots, d_{t,i}, \cdots, d_{t,I}\}$, とする．ここで I は時刻 t での文書数を表す．MDTM は時刻 t での各文書 $d_{t,i}$ が固有のトピック比率 $\boldsymbol{\theta}_{t,i} = \{\theta_{t,i,k}\}_{k=1}^{K}$ を持つとする．ある単語 $w_{t,j}$ は，潜在トピック $z_{t,k}$ を $\boldsymbol{\theta}_{t,i}$ に従って選択した後にトピックでの単語分布 $\boldsymbol{\phi}_{t,k} = \{\phi_{t,k,j}\}_{j=1}^{J}$ に従って生成を行っている．ここで単語 $w_{t,j} = \{w_{t,1}, \cdots, w_{t,J}\}$, J は単語数，潜在トピック $z_{t,k} = \{z_{t,1}, \cdots, z_{t,K}\}$, K は潜在トピック数と定義する．単語分布 $\boldsymbol{\phi}_{t,k}$ の生成をする際，多重スケールでの時間発展を考慮し，生成を行っている．ここで多重スケールを $s_l = \{s_1, \cdots, s_L\}$ とする．スケールとはトピックが持つ単語分布の時間単位であり，L はスケール数である．また s_l は時刻 $2^{l-1}+1$ 毎の t の時間幅を表しており，l が増えるほど長い期間である長期スケールになっていく．このスケールを考慮し，単語分布 $\boldsymbol{\phi}_{t,k}$ は時刻 $t-1$ における複数の時間スケールでの単語分布 $\{\hat{\boldsymbol{\omega}}_{t-1,k,l}\}_{l=1}^{L}$ を基に生成される．これによりスケール毎での依存性をモデルに組み込めるためモデルの頑健性を高めることができる．ここでスケール s_l の単語分布に与える重みを $\lambda_{t,k,l}$ としたとき，$\{\hat{\boldsymbol{\omega}}_{t-1,k,l}\}_{l=1}^{L}$ は単語分布 $\boldsymbol{\phi}_{t,k}$ の事前分布として，平均を多重スケール単語分布の重み付け和とする以下のディリクレ分布を用いる．

$$\boldsymbol{\phi}_{t,k} \sim Dirichlet(\sum_{l=0}^{L} \lambda_{t,k,l} \hat{\boldsymbol{\omega}}_{t-1,k,l}). \tag{1}$$

これにより時間スケールを考慮した単語分布 $\boldsymbol{\phi}_{t,k}$ が出力される．

2.2　多重スケール単語分布推定

式 (1) でのパラメータである多重スケール単語分布 $\omega_{t,k,l,j}$ の推定をしていく．$\omega_{t,k,l,j}$ は時刻 $t-2^{l-1}+1$ から t におけるトピック $z_{t,k}$ での単語 $w_{t,j}$ の出現確率を表している．そのため推定値は以下のようになる．

$$\begin{aligned}\hat{\omega}_{t,l,k,j} &= \frac{\sum_{i|d_i \in \boldsymbol{D}_t} tf_{i,j,k}}{\sum_{i|d_i \in \boldsymbol{D}_t} \sum_j tf_{i,j,k}} \\ &= \frac{\sum_{t'=t-2^{l-1}+1}^{t} \sum_{i|d_i \in \boldsymbol{D}_{t'}} tf_{i,j,k}}{\sum_{t'=t-2^{l-1}+1}^{t} \sum_{i|d_i \in \boldsymbol{D}_{t'}} \sum_j tf_{i,j,k}}\end{aligned} \tag{2}$$

$\sum_{i|d_i \in \boldsymbol{D}_t} tf_{i,j,k}$ は時刻 t における単語出現頻度を表している．

式 (1) で示した単語分布のディリクレ事前分布のパラメータは各時刻の単語分布の重み付き和として表現しているため式 (2) を用いて次のように得られる．

$$\begin{aligned}&\sum_{l=1}^{L} \lambda_{t,k,l} \hat{\omega}_{t-1,k,l,j} \\ =& \sum_{l=1}^{L} \lambda_{t,k,l} \frac{\sum_{t'=t-2^{l-1}+1}^{t-1} \sum_{i|d_i \in \boldsymbol{D}_{t'}} tf_{i,j,k}}{\sum_{t'=t-2^{l-1}+1}^{t-1} \sum_{i|d_i \in \boldsymbol{D}_{t'}} \sum_j tf_{i,j,k}} \quad .\end{aligned} \tag{3}$$

これにより前の時刻を事前分布としているため，MDTM は過去の時刻でのモデルを考慮した時系列モデルとなっている．

3　提案手法

従来手法では複数の時間スケールを仮定し，各スケール毎での単語分布を平滑化させ，各トピックでの単語分布としている．しかし，長期スケールでの単語分布にモデルが依存してしまうため，短期スケールで出現する話題の変化などによりトピックが変化した単語，例えば，長期スケールではスポーツトピックに属する単語が短期スケールにおいては政治トピックに変化するような単語を含む文書に対し，正しく単語分布が出力されない可能性がある．また，MDTM の出力である $\boldsymbol{\phi}_{t,k}$ は潜在トピックであるため，単に文書分類に適用することは困難である．そこで本提案では，トピックが変化する単語を含む文書に対しても分類精度を向上させるため，$\boldsymbol{\phi}_{t,k}$ の事前分布パラメータに対し，短期スケール，長期スケールでの事前分布パラメータを比較し，変化量が大きい場合に限り，$\boldsymbol{\phi}_{t,k}$ の事前分布パラメータを短期スケールの事前分布パラメータとすることを考える．また,MDTM の出力である $\boldsymbol{\phi}_{t,k}$ をナイーブベイズに用い，潜在トピックにカテゴリ情報を持たせることで文書分類に適用させる．

3.1 スケール変化における単語分布のパラメータの導出

まず，複数のスケールを短期スケール $S_a^{sh} = \{s_1, \cdots, s'_L\}$，長期スケール $S_b^{lo} = \{s_{L-L'}, \cdots, s_L\}$ と定める．ここで L' は短期スケールでの最大スケール数とする．式 (4) はスケール毎のディリクレ事前分布パラメータを表しており，ここでのスケール期間を変更することにより式 (1) での $\boldsymbol{\phi}_{t,k}$ の推定に影響してくる．短期スケールでのディリクレ事前分布パラメータを $\alpha_{t,k,j}$, 長期スケールでのディリクレ事前分布パラメータを $\beta_{t,k,j}$ としたとき，それぞれ式は次のようになる．

$$\alpha_{t,k,j} = \sum_{l=0}^{L'} \lambda_{t,k,l} \hat{\omega}_{t-1,k,l,j}. \tag{4}$$

$$\beta_{t,k,j} = \sum_{l=L-L'}^{L} \lambda_{t,k,l} \hat{\omega}_{t-1,k,l,j}. \tag{5}$$

また，本手法では短期スケールと長期スケールでの単語分布の変化量が大きい場合に限り，短期スケールを重視するため $\alpha_{t,k,j}$ と $\beta_{t,k,j}$ の変化量 $\frac{\alpha_{t,k,j}}{\beta_{t,k,j}}$ により短期スケールを重視するか決定される．よって全スケールにおいてのディリクレ事前分布パラメータ $\boldsymbol{\psi}_{t,k}$ は次のようになる．

$$\boldsymbol{\psi}_{t,k} = \begin{cases} \sum_{l=0}^{L} \lambda_{t,k,l} \hat{\omega}_{t-1,k,l,j} & \text{if } \frac{\alpha_{t,k,j}}{\beta_{t,k,j}} < \mu, \\ \frac{L}{L'} \left(\sum_{l=0}^{L'} \lambda_{t,k,l} \hat{\omega}_{t-1,,k,l,j} \right) & \text{otherwise.} \end{cases} \tag{6}$$

ある閾値 μ よりも単語分布の変化量が少ない場合，従来手法で行っていた全スケールでの単語分布を用いた，平滑化が行われる．また，単語分布の変化量が大きい場合，短期スケールでの単語分布のみを用い，ディリクレ事前分布パラメータとしている．

3.2 文書分類への適用

式 (1) で得られたトピック毎での単語分布 $\boldsymbol{\phi}_{t,k}$ から文書分類に適用する．しかし，$\boldsymbol{\phi}_{t,k}$ はカテゴリ情報を持っていないため，単に $\boldsymbol{\phi}_{t,k}$ を文書分類をすることは困難である．そこでカテゴリ毎に MDTM を学習させていき，カテゴリ毎の単語分布 $\boldsymbol{\phi}_{t,m,k}$ を導き，ナイーブベイズを用いることで文書分類に適用することを考える．ここでカテゴリを $c_m = \{c_1, \cdots, c_M\}$, M はカテゴリ数を表している．まず時刻 t におけるナイーブベイズは次のように示されている．

$$P(c_m|d_{t,i}) = \frac{P(c_m)P(d_{t,i}|c_m)}{P(d_{t,i})} \propto P(c_m)P(d_{t,i}|c_m). \tag{7}$$

ここから上式を解くために, まずトピック $z_{t,k}$ を用いて $P(d_{t,i}|c_m)$ を表すとまずトピック $z_{t,k}$ を用いて $P(d_{t,i}|c_m)$ を表すと

$$P(d_{t,i}|c_m) = \sum_k P(d_{t,i}|z_{t,k})P(z_{t,k}). \tag{8}$$

となる．ここで新規文書の単語頻度ベクトル $\boldsymbol{y} = \{y_1, \cdots, y_N\}$ とする.N は新規文書の単語数とする．$P(d_{t,i}|z_{t,k})$ を MDTM のカテゴリ毎の出力である $\phi_{t,k,j}$ を用いて表すと次のようになる．

$$P(d_{t,i}|z_{t,k}) = \sum_k \left(\prod_j \phi_{t,k,j} \right)^{y_j}. \tag{9}$$

これによりカテゴリ情報を持たすことが出来るため，MDTM の出力 $\boldsymbol{\phi}_{t,k}$ を文書分類に適用させている．

4 実験

提案手法の有効性を示すため，実際の時系列データである新聞記事データを用いて分類実験を行い，分類精度の評価を行った．

4.1 実験条件

実験には毎日新聞 2005 年 (1 月 1 日〜12 月 31 日) の 4 カテゴリ (スポーツ，経済，政治，芸能) の記事データを使用する．1 月から 11 月までの記事データからランダムに選ばれた 900 件 ×4 カテゴリを学習データとし，12 月の記事データからランダムに選ばれた 60 件 ×4 カテゴリをテストデータとし，3 つのデータセットを用いる．ただし，データセットは日付の連続した記事データとなっている．また，時間単位 t=1 日としている．スケール数は最大スケールの分布がデータ全期間を含むように $L = [\log T + 1]$ と設定した．ここで T は時刻数である．提案手法で用いられる短期スケール S_a^{short} の最大スケール数 L'=5, 変化量の閾値 μ=0.5 とし，実験を行った．

4.2 実験結果と考察

従来手法，提案手法の時刻 $t = 1$，$t = 7$ においての実験結果を図 1 に示す．

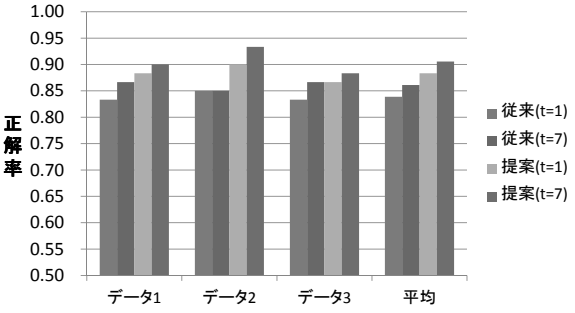

図 1．実験結果

図 1 よりすべてのデータに対し，提案手法の正解率が勝っていることから，その有効性を示すことができた．

提案手法が従来手法よりも良い結果を示したのは，短期スケールにおいての事前分布パラメータの変化量によりパラメータの選択を行っているため，時系列データにおける話題の変化，単語のトピック変化に対応したためだと考えられる．また，時刻 $t = 1$ での結果が悪くなったのは，時刻 $t = 1$ においてはまだ学習データが少なく，スケールも考慮出来ないためだと考えられる．また時刻 $t = 7$ において精度が上がっていることから過去のモデルの依存性とスケールによる効果が大きいと考えられる．

5 まとめと今後の課題

本研究では，MDTM を拡張し，短期スケールの事前分布パラメータの変化量が大きい場合，トピック毎の単語分布の事前分布パラメータとすることで時系列変化によってトピックが変化した単語にも対応した手法を提案し，実験によりその有効性を示した．

今後の課題としては，短期スケールでの最大スケール数 S' や閾値 μ の自動設定などが挙げられる．

参考文献

[1] 岩田具治, 山田武士, 櫻井保志, 上田修功,“ オンライン学習可能な多重スケールでの時間発展を考慮したトピックモデル．”情報論的学習理論テクニカルレポート,2009.