時間的尺度を導入した主要潜在トピック抽出法

1G06H044-5　小部泰嗣
指導教員　後藤正幸

1　研究目的

近年，World Wide Web は，コミュニケーションの新たなメディアとして発展し続けており，日々膨大な数の電子文書が公開されている．このような電子文書は時間と共に変化する流行や時事問題といったトピックを有している．そして，配信される多くの電子文書に，公開日時等の時間情報が付与されていることから，この時間情報を分析に取り入れ，ある時期に話題となったテーマなどを自動抽出できれば，有用な知識が得られると考えられる [1], [2].

文書に時間情報が付与された文書群を文書ストリームと呼ぶ．単なる話題抽出だけでなく，ある時期における主要な出来事, 話題, 技術などのテーマをトピックとして抽出し, 文書ストリームの全貌を理解することを目的とした手法の 1 つとして PMM(Parametric Mixture Models) 型主成分分析法 [1] がある．この手法は，長期間の文書ストリームを理解するために効果的であることが示されている．しかし，短期間で変化するトピックを抽出しようとした場合，時間的相関が構造的相関に埋もれてしまい，うまく主要トピックを抽出できないという問題がある．

そこで本研究では，PMM 型主成分分析法が短期間の文書ストリームにも効果的となるために単語の特徴量算出法として，新たに IAT (Inverse of Appeared Times) という尺度を提案し，この特徴量を用いたトピック抽出法を示す．また，提案手法を短期間における毎日新聞の国際面記事への分析に適用し, 有効性を示す.

2　PMM 型主成分分析法 [1]

PMM は，上田ら [2] によりモデル化された多重トピックを有するテキストの確率モデルである．

2.1　文書ストリームの確率的生成モデル

文書ストリーム $\mathcal{D}$ において，文書ストリームの時間ステップを $t(1 \leq t \leq T)$ とする．また $\mathcal{D}$ における単語 $w_i(1 \leq i \leq V)$ の集合を $\boldsymbol{\mathcal{W}} = \{w_1, \cdots, w_V\}$ とする．時間ステップ t における文書群を $\boldsymbol{D}(t)$，文書の総数を $N(t)$，第 n 番目の文書を $d(t,n)$，文書 $d(t,n)$ の単語頻度ベクトルを $\boldsymbol{x}(t,n) = (x_1(t,n), \cdots, x_V(t,n))$ とし，$\boldsymbol{D}(t)$，$d(t,n)$ 内の単語の総頻度数をそれぞれ $M(t)$，$M(t,n)$ と表す.

このとき，時間ステップ t での単語 w_i の生起確率 $\psi_i(t)$ を，

$$\psi_i(t) = \left(1 - \sum_{l=1}^{L} h_l(t)\right)\overline{\psi}_i + \sum_{l=1}^{L} h_l(t)\phi_{li}, \tag{1}$$

とする．ここに，文書ストリーム $\mathcal{D}$ には，主要潜在トピック $l(1 \leq l \leq L)$ とともに，1 つの通常トピックの存在を仮定する．ϕ_{li} は，主要潜在トピック l の単語 w_i の生起確率であり，$\overline{\psi}_i$ は通常トピックの単語 w_i の生起確率である．主要潜在トピック l の主要アクティブ期間を $[s_l, e_l]$ とし，

$$h_l(t) = \begin{cases} c_l, & \forall t \in [s_l, e_l]; \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

と仮定する．ただし，主要アクティブ期間とはその潜在トピックが顕著に存在する期間のことで，各 c_l は，$0 < c_l \leq 1$, $\sum_{l=1}^{L} c_l \leq 1$ となる定数である．

2.2　トピック軸の推定法

主成分分析で抽出される主成分をトピック軸として推定する．まず，各 l に対して，第 l トピック軸を推定することを考える．推定されるトピック軸は，単体 $\triangle^{V-1}$ 上でサンプルデータ $\boldsymbol{D}$ の射影値 $A_0(t;\boldsymbol{u})$ の関数 $F(\boldsymbol{u})$ を最大にする単位ベクトル $\boldsymbol{u}$ で与えられる．

ただし，$\boldsymbol{u}$ によって与えられる射影値 $A_0(t;\boldsymbol{u})$ と関数 $F(\boldsymbol{u})$ は，

$$A_0(t;\boldsymbol{u}) = A(t;\boldsymbol{u}) - \overline{\boldsymbol{\theta}} \cdot \boldsymbol{u}, \tag{3}$$

$$F(\boldsymbol{u}) = \sum_{t=1}^{T} M(t)\Big\{\Big(\frac{1}{M(t)}\boldsymbol{X}(t) - \overline{\boldsymbol{\theta}}\Big) \cdot \boldsymbol{u}\Big\}^2, \tag{4}$$

で与えられ，

$$A(t;\boldsymbol{u}) = \frac{1}{M(t)}\boldsymbol{X} \cdot \boldsymbol{u}, \tag{5}$$

$$\overline{\theta}_i = \frac{\sum_{t=1}^{T} X_i(t)}{\sum_{t=1}^{T} M(t)} = \frac{\sum_{t,n} x_i(t,n)}{\sum_{i,t,n} x_i(t,n)}, \tag{6}$$

であり，さらに, $\overline{\boldsymbol{\theta}} = (\overline{\theta}_1, \cdots, \overline{\theta}_V)$ である．また，$\triangle^{V-1}$ は，V 次元ユークリッド空間 R^V の $(V-1)$ 次元標準単体で，$\triangle^{V-1}$ 上でのサンプルデータは，$\boldsymbol{D} = \frac{1}{M(t)}\boldsymbol{X}(t)$ である．さらに，文書 $\boldsymbol{D}(t)$ の単語頻度ベクトルは $\boldsymbol{X}(t) = (X_1(t), \cdots, X_V(t))$ であり，このとき $X_i(t) = \sum_{n=1}^{N(t)} x_i(t,n)$ で与えられる．

Lagrange 乗数法により，$\boldsymbol{u}_l(1 \leq l \leq L)$ は，次式を要素とする $V \times V$ 実対称行列 $B = [b_{i,j}], (1 \leq j \leq V)$ の長さ 1 の第 l 固有ベクトルによって求められる．

$$b_{i,j} = \sum_{t=1}^{T} M(t)\Big(\frac{X_i(t)}{M(t)} - \overline{\theta}_i\Big)\Big(\frac{X_j(t)}{M(t)} - \overline{\theta}_j\Big). \tag{7}$$

2.3　主要アクティブ期間の推定法

t を固定して考えた場合，確率変数 $A(t;\boldsymbol{u})$ は中心極限定理より，近似的に平均 $\mu(t;\boldsymbol{u})$，分散 $\sigma(t;\boldsymbol{u})^2/M(t)$ のガウス分布 $N(\mu(t;\boldsymbol{u}), \sigma(t;\boldsymbol{u})^2/M(t))$ に従うと近似できる．

t を変数としてみた場合，確率変数 $A(t;\boldsymbol{u}_l)$ は t がアクティブ期間のときとそれ以外のときで異なるガウス分布に従い，

$$A(t;\boldsymbol{u}_l) \sim \begin{cases} N(\mu_l, {\sigma_l}^2/M(t)), & \forall t \in [s_l, e_l]; \\ N(f_l, {g_l}^2/M(t)), & \text{otherwise}, \end{cases} \tag{8}$$

のように近似できる．式 (8) のモデルを用いて，$\mu_l, {\sigma_l}^2, f_l, {g_l}^2$ の最尤推定値は s_l と e_l の全探索で求まる [1].

2 クラス分類に対する最尤推定法より，各クラスの条件付き確率密度をガウス分布とし，t がアクティブ期間のとき $a_t = 1$，それ以外のとき $a_t = 0$，さらに $\alpha = e_l - s_l + 1$ とすると，第 l 主成分の尤度関数 $\mathcal{L}_l$ は次式のようになる．

$$\mathcal{L}_l = \prod_{t=1}^{T}(\alpha\mathcal{N}(M(t) \mid \mu_l, {\sigma_l}^2/M(t)))^{a_t} \times\{(T-\alpha)\mathcal{N}(M(t) \mid f_l, {g_l}^2/M(t))\}^{1-a_t}. \tag{9}$$

この式 (9) に対数をとり，s_l と e_l の全探索を行うことで，対数尤度関数を最大とするアクティブ期間 $[s_l, e_l]$ が求まる．ただし，$\mathcal{N}(x \mid \mu, \sigma^2)$ は平均 μ，分散 ${\sigma_l}^2$ の正規分布に対し，データ x が与えられた場合の尤度である．

2.4 トピック文書のランキング

各 $l(1 \leq l \leq L)$ に対し，推定したトピック軸と主要アクティブ期間に基づいて，第 l 主要潜在トピックを表す文書群を抽出する．文書 $d(t,n)$ の重要度を判定する第 l トピック度 $r_l(d(t,n))$ を次のように定める．

$$r_l(d(t,n)) = \frac{A_l(t,n;\boldsymbol{v}_l)}{\overline{\sigma}_l/\sqrt{M(t,n)}}. \tag{10}$$

ただし，$A_l(t,n;\boldsymbol{v}_l)$ は文書 $d(t,n)$ に対して，正規化データ $\frac{1}{M(t,n)}\boldsymbol{x}(t,n) \in \triangle^{V-1}$ の第 l トピック軸への射影値であり，$\boldsymbol{v}_l$ は，$\mu_l \geq f_l$ のとき $\boldsymbol{u}_l$ を，それ以外のとき $-\boldsymbol{u}_l$ を取る単位ベクトルである．$A_l(t,n;\boldsymbol{v}_l)$ は，ガウス分布に従う確率変数と仮定でき，$\overline{\sigma}_l$ は最尤推定法により求まる．

3 提案手法

3.1 準備

従来手法 [1] では，長期間のデータに対して手法を適用している．しかし，短期間の文書ストリームに対して，ミクロな視点でのトピック抽出を行おうとする場合，時間的なトピック抽出ができないという問題がある．これは時間経過によって単語総数 $M(t)$ の変化が長期間では顕著であるが，短期間では微少である為に，時間的相関でなく構造的相関を抽出してしまうことに起因する．そこで本研究では，構造的相関に埋もれた時間的相関の特徴を強調して取り出すため，単語に対して時間的重要度による重みを付与する方法を提案する．

3.2 IAT の定式化

各単語 w_i において，何期間出現したかを表す値を $at(i)$ とする．このとき，時間ステップ t で，単語 w_i が存在するとき $\beta(i,t) = 1$，それ以外のとき $\beta(i,t) = 0$ とすると，$at(i)$ は，

$$at(i) = \sum_{t=1}^{T} \beta(i,t), \tag{11}$$

となる．さらに，少ない期間に集中して出現した単語に高い重要度を付与するため，$at(i)$ の逆数をとり，各単語 w_i の時間的な重みを，

$$iat(i) = \log \frac{T}{at(i)}, \tag{12}$$

で与えるものとする．これを Inverse of Appeared Times (IAT) 尺度と呼ぶ．$iat(i) \cdot x_i(t,n)$ を要素とする V 次元ベクトルを $\boldsymbol{x}^*(t,n)$ とし，これを $\boldsymbol{x}(t,n)$ と置き換えて 2 節の PMM 型主成分分析法を実行し，トピック軸を推定する．

4 評価実験

提案手法の有効性を検討するために実データを用いて評価実験を行った．

4.1 実験条件

実験データは毎日新聞 2000 年の国際面記事を用いる．本データセットにおける総文書数は 9025 であり，形態素解析後の語彙総数 $V = 8173$ となった．ここで，時間ステップは 1 日とし，$T = 365$ とする．

4.2 実験結果

提案手法，従来手法の第 1 トピック上位 10 文書を図 1 に示す．それぞれの第 1 トピックの主要アクティブ期間は，提案手法は 2/14〜3/13，従来手法は 4/9〜4/10 となった．

＜提案手法の第1トピック上位10位文書タイトル＞

日付	タイトル
2/29	共和党の候補争い，正念場－3月7日は「スーパーチューズデー」
3/09	マケイン現象とは何だったのか？
2/24	ミシガン・アリゾナ州　マケイン氏が勝利，ブッシュ氏に打撃
3/07	共和党候補氏名争い　天王山…今日スーパーチューズデー
3/02	マケイン現象/下 漁夫の利ゴア氏 無党派，本選では敵にも
2/29	マケイン現象/上「権威に挑戦」が奏功
3/07	今日スーパーチューズデー NY州の勝敗カギ
2/24	マケイン氏なぜ強い－好印象，無党派つかむ
3/08	マケイン旋風に苦戦－共和党，ブッシュ氏指名へ
3/08	激戦区ルポ スーパーチューズデー，共和党に深い亀裂

＜従来手法の第1トピック上位10位文書タイトル＞

日付	タイトル
4/10	総選挙と経済支援、「南北」の利害一致
4/9	森に消えたカルト・ウガンダ集団死事件／下
4/10	フジモリ氏とトレド氏、決選投票の公算大
4/9	国民大会の「非常設化」、微妙に　親民党など反対の構え
4/9	「司法改革が次期政権の焦点」－－フジモリ・ペルー大統領
4/10	北京で秘密接触　金大統領、側近を派遣
4/10	京都議定書、発効　両論併記の宣言採択
4/10	米、CIA職員を処分　－－中国大使館の誤爆事件
4/9	[韓国総選挙]落選運動も「最後の訴え」
4/10	南北首脳会談　合意書の全文

図 1. 提案手法と従来手法の第 1 トピック上位 10 文書

この結果，従来手法では明らかに内容の異なる記事が同一トピックとして得られているのに対し，提案手法では同じ内容の記事が同一トピックとして得られていることが分かる．第 2 トピック以下も同様の結果となった．

4.3 まとめ及び考察

1. 図 1 より，提案手法が従来手法よりも効果的なトピック抽出を行えていることがわかる．すなわち，提案手法により，主要トピックと関連する文書群を抽出することが可能となったといえる．
2. IAT 尺度により，全ての期間で出現する一般的な単語は重要度が低くなり，逆に短期間で出現が集中する特徴的な単語の重要度が上がることで，構造的相関の中に埋もれていた時間的相関の特徴が強調される．これにより，比較的短期間データに対しても，PMM 型主成分分析法が時間的相関を抽出できたといえる．

5 まとめと今後の課題

本研究では，トピック抽出を目的とした PMM 型主成分分析法への IAT 尺度の導入を提案し，実験により提案手法の有効性を示した．

今回はデータとして新聞記事を用いたが，さらに World Wide Web 上の文書ストリームであるブログ，掲示板，クチコミなどへの適用が今後の課題である．

参考文献

[1] 木村晶弘, 斉藤和巳, “PMM 型主成分分析を用いた文書ストリームの主要潜在トピック 抽出, ” 日本応用数理学会論文誌, Vol. 18, No. 3, pp. 363–388, 2008.

[2] 上田修功, 斉藤和巳, “多重トピックテキストの確率モデル－パラメトリック混合モデル－, ” 電子情報通信学会論文誌 D-Ⅱ, Vol. J87-D-Ⅱ, pp. 872–833, March 2004.