Vol.2009-MPS-76 No.51
Vol.2009-BIO-19 No.51
2009/12/18

文書ストリームからのホットトピック抽出を目的としたSR法の拡張

東野正行[†1]　熊野雅仁[†2]
木村昌弘[†2]　斉藤和巳[†3]

文書ストリームからホットトピック文書群を抽出する手法として，ネットワークコア抽出法である SR 法を拡張した手法を提案する．新聞記事ストリームデータ及び人工文書ストリームデータを用いた実験により，提案法は従来法よりも高精度であることを示す．

Extending SR-method for Extracting Hot Topics in Doucument Streams

Mayuki Higashino,[†1] Masahito Kumano,[†2]
Masahiro Kimura[†2] and Kazumi Saito[†3]

We propose a method for extracting hot-topic documents in a document stream. The proposed method extends the SR-method for network-core extraction. Using real and synthetic document stream data, we experimentally demonstrate that the proposed method outperforms convetional methods.

†1 龍谷大学大学院理工学研究科電子情報学専攻
Divsion of Electronics and Informatics, Graduate School of Science and Technology, Ryukoku University
†2 龍谷大学理工学部電子情報学科
Department of Electronics and Infromatics, Faculty of Science and Technology, Ryukoku University
†3 静岡県立大学経営情報学部
School of Administration and Informatics, University of Shizuoka

1. はじめに

インターネットの普及・発展に応じて，様々なメディアで配信される膨大な情報がネットワーク内に取り込まれ，蓄積され続けており，インターネットは情報の宝庫となっている．このような膨大な情報群から，必要な情報を自動的に得るためのコンテンツ分析技術が注目されている．Shao ら[4] は，コンテンツ分析技術の一領域として，電子メールデータのような送信者と受信者などの社会的関係性情報をも含む特殊な時系列性を持つ文書ストリームデータに対して，文書ネットワークを含む複数のネットワークを分析し，それらの結果を統合することにより，主要なイベントを精度良く抽出する手法を提案している．

一方，コンテンツ分析の領域として，トピック分析がある．トピック分析には，活性度の高い状態にあるトピックをホットトピックと呼び，ホットトピックを効率良く抽出する研究がある．Kleinberg[1] は，バースト性に基づいて文書ストリームのホットトピック抽出を行う手法を提案している．本稿では，新聞記事などの大規模文書ストリームを対象としたホットトピックの抽出法に焦点を当てる．また，トピックは，複数のトピックが独立して存在する場合よりも，多重のトピックが混在する多重トピック性を持つ場合が多く，そのような特徴を持つ文書群から効率良くホットトピックを抽出する手法が望まれる．

ところで，Saito ら[2] は，グラフ上のリンクが密結合するコア部を効率良く抽出する数理的手法の一つとして SR(Spectral Relaxation) 法を提案している．また，Saito らは，インターネット上のブログのトラックバックネットワークを対象として，SR 法を適用し，抽出されたコア部の主要トピックを分析する問題に応用し，成果を上げている．SR 法はネットワークにおけるノードの多重抽出を可能としているため，文書ストリームにおけるトピックの多重性にも対応できる可能性がある．しかし，文書ストリーム中の文書どうしはリンクで結ばれていないため，SR 法は，文書ストリームには直接的に適用できない．一方，文書ストリームは，各文書内の単語情報に基づいた文書間類似度を計算し，類似度を重みとするリンクを文書間で結ぶことにより，文書ネットワークを構築することができる．本稿では，文書ネットワークを対象とし，文書間の類似度と時間情報を扱えるよう SR 法を拡張することで，拡張 SR 法が文書ストリームのホットトピック抽出問題に適用できることを示す．

大規模人工文書ストリームデータおよび大規模実文書ストリームデータを用いた二つの実験により，提案法の特徴を示し，主要なホットトピック抽出法との比較実験により，提案法の有効性を示す．

Vol.2009-MPS-76 No.51
Vol.2009-BIO-19 No.51
2009/12/18

2. SR 法

SR 法とは SEO スパム検出において優れた性能を示した主成分ベクトル法を一般化した手法である．基本的なアイディアとしては，ネットワークの比較的リンクが密集するコア部には潜在トピックが内在していると考え, ネットワークの全てのノード集合における平均リンク数が最大となるノード集合を探索することで，コア部の発見を行い，ホットトピックの抽出を試みるというものである．

ここで，ネットワークの全ノード集合を $S = \{1, ...N\}$ とし，その隣接行列を A とする．すなわち隣接行列の第 (i, j) 成分 $A(i, j)$ は，ノード i と j 間にリンクがある場合は 1 に，なければ 0 に設定する．簡略化のため，自己リンクはないものとし，リンクは無方向であるものとする．すなわち $A(i, i) = 0$ かつ $A(i, j) = A(j, i)$ である．また，ノード集合 $C \subset S$ に対し，その平均リンク重みは以下の式で定義することができる．

$$G(C) = \left(\sum_{i \in C} \sum_{j \in C} (A_{(i,j)} / |C|) \right) / 2$$

$|C|$ はノード集合の要素数である．既に述べたように，潜在トピックがあれば平均リンク数の高いノード集合が形成されると仮定する．そこで $G(C)$ を最大にするノード集合 C の探索問題を考える．しかし単純な数え上げによる網羅的探索ではノード数の多い大規模ネットワークにて組み合わせ爆発が起きる．よって SR 法では緩和問題が最適に解けることに着目したアプローチを採用している．

ノード集合 C に対して N 次元ベクトル $\mathbf{q}$ を，$i \in C$ なら $q_i = 1$，それ以外なら $q_i = 0$ と定義する．この時，平均リンク数の定義は

$$G(\mathbf{q}) = \mathbf{q}^T \mathbf{A} \mathbf{q} / 2 \mathbf{q}^T \mathbf{q}$$

と書き換えることができる．ここでベクトル $\mathbf{q}$ の各要素に対して連続値まで許容すれば，この式の右辺は Rayleigh 商に他ならない．よって $G(\mathbf{q})$ の最大値は，行列 A にて固有値を最大にする固有ベクトル $\mathbf{q}^*$ で与えられる．固有ベクトル $\mathbf{q}^*$ を求めるには，以下のパワー法を土台としたアルゴリズムが適応できる．

E1 $t = 1, \mathbf{q}^{(0)} = (1, ..., 1)^T$ と初期化する．

E2 $\tilde{\mathbf{q}} = \mathbf{A}\mathbf{q}^{(t-1)}, \mathbf{q}^{(t)} = \tilde{\mathbf{q}} / max_i \tilde{q}(i)$ を求める．

E3 $max_i |q^{(t)}(i) - q^{(t-1)}(i)| < \varepsilon$ なら反復を終了する．

E4 $t = t + 1$ として **E2** に戻る．

ここで ε は終了条件を制御する正の実数であり，反復終了後に $\mathbf{q}^* = \mathbf{q}\ (t)$ として結果が求まる．上記アルゴリズムで固有値最大の固有ベクトルが基まり，基本設定の妥当な緩和問題を解いていると言える．

固有ベクトル $\mathbf{q}^*$ の各要素をバイナリ化することで基本問題の解を求める．まず $\mathbf{q}^*$ の要素の大小に基づき各ノードをランキングすればリスト $R = [r(1), ...r(N)]$ が求まる．なお tie-break は任意に行うとする．$R = [r(1), ...r(N)]$ の上位 k 個のノード集合

$$C(k) = \{r(i) : i \leq k\}$$

を考えればその平均リンク数は以下の式で求まる．

$$G(k) = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} A\left(r(i), r(j)\right) / k$$

この式を最大にする k^* を探索してノード集合 $C(k^*)$ を求める．

効率良く k^* を探索するために以下の漸化式を利用する．

$$G(k+1) = G(k) + (\Delta(k+1) - G(k))/k + 1$$

ここで $\Delta(k+1)$ はノード $r(k+1)$ を加えたことによるリンク数の増分であり以下で計算できる．

$$\Delta(k+1) = \sum_{j=1}^{k} A(r(j), r(k+1))$$

一方，定義より $G(1) = 0$ である．手順をまとめれば以下となる．

F1 $\mathbf{q}^*$ の要素をソートしランク関数 $r(i)$ を求める

E2 漸化式から $G(2), ...G(N)$ を求める．

E3 $k^* = argmax_k G(k)$ を求めて $C(k^*)$ を出力する．

上述した処理手順を M 回繰り返して実行することにより結合が密な M 個のコア部を抽出できる．すなわちアルゴリズムは以下となる．

G1 $m = 1$ から M までの以下のステップを実行する．

G2 **E1** から **E4** を反復させ $\mathbf{q}^*_{\mathbf{m}}$ を求める．

G3 **F1** から **F3** で $C_m(k^*)$ を求める．

G4 $i, k \in C_m(k^*)$ ならば $A(i, j) = 0$ に設定する．

最終的に，結果は $C^*_1, ...C^*_M$ で求まる．**G4** の操作によって消去しているのは抽出するコア内のリンクのみで，抽出したノード自体は消去していないため，同一ノードが別のコア部

Vol.2009-MPS-76 No.51
Vol.2009-BIO-19 No.51
2009/12/18

に属するノードとして複数回抽出される可能性があり，文書の多重トピック性を考慮していると言える．

ところで，SR 法では，ある瞬間のネットワークの構造のみを考慮して，そのネットワークのコア部を抽出するため，ノードやリンクの発生時期は基本的に考慮されていない．そのため，例えば，あるイベントが発生してから期間が空いた後に，最初に発生したイベントと類似したイベントが発生した場合，本来ならば期間が空いているために異なるトピックのイベントであると位置づけたいが，時系列データとして期間を考慮していないために，同一トピックのイベントと見なされ，同じコア部としてひとまとめに抽出される可能性がある．そのため，時系列データの期間情報を扱うホットトピック抽出問題の場合，SR 法を直接適応するのには問題があると考えられる．

3. 提 案 法

本研究では，文書ストリームデータから文書間の類似性に基づいて文書ネットワークを構築し，文書ネットワークからホットトピックを効率良く抽出するため，期間情報を扱えるよう拡張した，拡張 SR 法を提案する．

3.1 文書ストリーム

本研究で対象とする文書ストリームデータは BOW（bag-of-wors）表現により表現されているものとし，また各文書には出現した順番に ID 番号が割り振られているものとする．すなわち t 日目に出現した ID 番号 n の文書を $d_{t,n}$ とすると，その文書における単語 i の出現頻度数 x_i を用いて文書ストリームデータは $d_{t,n} = \{x_i; i = 1, ...V\}$ と表される．また，文書ストリームデータ全体は $D = \{d_{t,n}; t = 1, ...T, n = 1, ...N\}$ と表現できる．

抽出されたホットトピック群 $C_k \subset D(I_k)$ は，文書群，及び文書群と共起性が強い期間 $I_k \subset (0, T)$ との組み合わせ集合を意味する．また，ホットトピックは，短期間に活性化状態にあるという性質を考え，$|I_k| \ll T$ とする．

3.2 文書ネットワーク

文書ストリームデータに提案法を適応させるためには，文書ストリームデータから文書ネットワークを構築する必要がある．以下に，文書ネットワークの構築法を述べる．

3.2.1 特 徴 量

まず，文書ストリームデータから単語を抽出し，BOW 表現から $TFIDF$ 値を求める．$TFIDF$ とは，ある文書を特徴づける単語を抽出するための指標であり，単語 i の $TFIDF$ 値を以下の式で求めることができる．

$$TFIDF_i = TF_i * IDF_i$$

$$TF_i = w_{n,i} / \sum_k w_{n,k}$$

$$IDF_i = \log N / |d : d \ni c_i|$$

$|d : d \ni c_i|$ は，単語 i を含む文書数である．上式の通り，$TFIDF$ 値は，単語の出現頻度に基づく TF と，逆出現頻度に基づく IDF による二つの指標の積で計算される．

IDF により，多くの文書に出現するいわゆる一般語の $TFIDF$ 値は下がり，注目する文書にしか出現しにくい単語の $TFIDF$ 値を上げることが可能となることから，$TFIDF$ 値は，ある文書で出現頻度が高く，かつ，その文書しか出現しにくい特徴的な単語を抽出することのできる良い指標となる．文書ネットワークでは文書をノードと捉え，文書と文書を繋げるリンクとして，異なる 2 つの文書同士の繋がりの強さを表す指標，すなわち文書間類似度を定義する必要がある．

本研究においては，文書内容と文書出現期間から文書間類似度を定義する．つまり，文書内容がより近い，また文書出現期間がより近い二つの文書同士の文書間類似度が高くなるよう，文書間類似度を定義した．

3.2.2 ネットワーク構築

まず，コサイン類似度に基づく文書内容類似度を定義する．文書ベクトル d_{n_1} と d_{n_2} の文書内容類似度 $docsim(d_{n_1}, d_{n_2})$ を以下の式として定義する．

$$docsim(d_{n_1}, d_{n_2}) = \overrightarrow{d_{n_1}} \cdot \overrightarrow{d_{n_2}} / |\overrightarrow{d_{n_1}}||\overrightarrow{d_{n_2}}|$$

次に，文書ベクトル d_{n_1} と d_{n_2} と，出現した日をそれぞれ $t_{d_{n_1}}$，$t_{d_{n_2}}$ とし，二つの文書の出現期間の日差と全期間の日数の比から，時間類似度 $timsim(d_{n_1}, d_{n_2})$ を以下の式で定義する．

$$timsim(d_{n_1}, d_{n_2}) = log(1 - |t_{d_{n_1}} - t_{d_{n_2}}|/T)$$

以上の定義に基づき，文書内容類似度と時間類似度の合計値を文書間類似度 $sim(d_{n_1}, d_{n_2})$ とする．これが構築したネットワークのノード間の関係性を表す指標となる．

$$sim(d_{n_1}, d_{n_2}) = docsim(d_{n_1}, d_{n_2}) + timsim(d_{n_1}, d_{n_2})$$

ここで，類似度に閾値 τ を設定し，τ 以下の類似度は 0 として文書ノード間のリンクを削除する．本稿の実験において閾値 τ は, 例えば二つの文書の文書内容が 80 %類似していても，それらの文書出現期間が全期間の 1/6 の期間離れていた場合，それら二つの文書は同じトピックに属する文書ではない, と判断できるような値に設定した．

Vol.2009-MPS-76 No.51
Vol.2009-BIO-19 No.51
2009/12/18

以上の手法により，本稿の実験では，各文書をノード，文書間類似度をリンク重みとして，文書ネットワークを構築した．

3.3 ホットトピック抽出

構築された文書ネットワークから，提案法のコア抽出法によりホットトピックを抽出する．ホットトピック群の抽出数を K と定めた時，抽出のアルゴリズムは以下となる．

(1) 文書ストリームデータから単語を抽出し $TFIDF$ 値を求める．
(2) 各文書間の文書間類似度を求め，文書ネットワークを構築する．
(3) 提案法により，文書ネットワークから重複を許してコア部を順次抽出し，K 個の文書群，すなわちホットトピック群 $\{ C_k; k = 1, \cdots, K \}$ を抽出する．

4. 評 価 法

4.1 評価尺度

各実験において，ホットトピックの抽出性能を評価するため，情報検索システムの性能評価に用いられる F 値に基づいた評価尺度を用いる．真のトピック l が L 個あるとき，真のトピック群を $\{H_l; 1 \leq l \leq L\}$ とする．ただし，それぞれのトピックは，各トピックごとに異なる文書数を持つ文書群であり，$|H_l|$ は，真のトピック l の文書数を表す．また，実験により，抽出されたトピックをトピック k としたとき，$|C_k|$ は，トピック k の文書数を表し，抽出された K 個のトピック群を $\{C_k; 1 \leq k \leq K\}$ とする．ただし，k は，$(L \leq k \leq K)$ とした．このとき，真のトピック l と抽出されたトピック k との F 値 $F_{l,k}$ は，

$$F_{l,k} = 2\,|H_l \cap C_k|\,/(|H_l| + |C_k|)$$

と表される．また，真のトピック l の L 個分の F 値の平均を取るため，平均 F 値 $F(K)$ を以下のように定義する．

$$F(K) = (\sum_{l=1}^{L} F_{l,k_l^*})/L$$

ただし，k_l^* は，真のトピック l に対し，抽出されたトピックの中で，最も良い F 値を持つトピック k で評価を行うため，$k_l^* = \arg\max_k F_{l,k}$ と定義される．これは，抽出したトピックと一致度の高い文書群をいかに抽出しているかに注目して評価を行うためである．

4.2 比 較 法

提案法の性能を評価するため，文書のバースト性を利用してホットトピックの抽出を行う Kleinberg 法，また，文書と期間の独立性からホットトピック抽出を試みる Fisher 検定法を提案法の性能を評価するための比較法として採用した．以下に，それぞれの手法の特徴を述べる．

4.2.1 Kleinberg 法

Kleinberg 法は，文書出現のバースト性に基づいてホットトピックを抽出する手法である．全期間で文書が一様に出現すると仮定した通常状態時の文書出現確率を $p_{k,0}$，あるトピック k に関する文書が生成されやすいバースト状態においての文書出現確率を $p_{k,1}$ とし，それぞれの二項分布尤度比でバースト度を定義する．

各時刻 t に出現した総文書数 N_t に対し，トピック k に関連した文書数が $n_{k,t}$ となる確率について，通常状態の場合の確率を $P_N(n_{k,t}; p_{k,0})$ とし，バースト状態の確率を $P_N(n_{k,t}; p_{k,1})$ とする．このとき，トピック k の期間 $[t_{i_1}, t_{i_2}]$ におけるバースト度は以下の式で求められる．

$$F(t_{i_1}, t_{i_2}; k) = \prod_{t=t_{i_1}}^{t_2} P_{N_t}(n_{k,t}; p_{k,1}) / \prod_{t=t_{i_1}}^{t_2} P_{N_t}(n_{k,t}; p_{k,0}) \quad (1)$$

Kleinberg 法では，単語 w_k を含む文書群をトピック k に関する文書群候補と考え，各文書群候補のバースト度が最大となるホットトピック期間 $[t_{i_1}, t_{i_2}]$ を求める．バースト度の大きい順に単語 w_k をランキングし，ホットトピック期間内で単語 w_k を含む文書群をホットトピックとして抽出する．

4.2.2 Fisher 検定法

Fisher 検定法は，2 つの事象が生起する際の独立性を検定する手法の一つである．ホットトピックを抽出する際には，単語と期間の独立性を検定することにより，ある期間と，その期間と関わりが深い単語を含む文書を抽出することができ，それらがホットトピック期間及びホットトピック文書群であると見なす手法である．まず，評価データにおける期間 I_i と単語 w_j に対して，以下のような 2 × 2 の分割表を用意する．

表 1 2 × 2 分割表
Table 1 2 × 2 contingency table

	w_j	$\overline{w_j}$	
I_i	a	b	m_i
$\overline{I_i}$	c	d	$N - m_i$
	n_j	$N - n_j$	

ここで，m_i は期間 I_i 内の全記事数，n_j は全記事中で単語 w_j を含む記事数を表す．a は期間 I_i 内の記事中に単語 w_j を含む記事数，b は期間 I_i 内の記事中に単語 w_j を含まない

Vol.2009-MPS-76 No.51
Vol.2009-BIO-19 No.51
2009/12/18

記事数，c は期間 I_i 以外の記事中に単語 w_j を含む記事数，b は期間 I_i 以外の記事中に単語 w_j を含まない記事数をそれぞれ表している．

以上の値を用いることで，検定値 F_i, j は，以下の式で求めることが出来る．

$$F_{i,j} = \sum_{k=a}^{min(m_i,n_j)} \left(\begin{array}{c} m_i \\ k \end{array} \right) \left(\begin{array}{c} N - m_i \\ n_j - k \end{array} \right) / \left(\begin{array}{c} N \\ m_i \end{array} \right)$$

上記の式によって求めた検定値 F_i, j に基づいて独立性の判定を行う．検定値 F_i, j の値が小さいと期間 I_i と単語 w_j は互いに独立であるという仮定は棄却され，共起性高いペアであるということになる．Kleinberg 法と同様に単語 w_j を含む文書群の検定値が最小となる期間 I_i を求め，検定値の小さい順に単語 w_j をランキングし，期間 I_i 内で単語 w_j を含む文書群をホットトピックとして抽出する．

5. 実データを用いた実験

提案法の有効性を検証するため，大規模データを用いて実験を行った．まず，使用した実データ及び提案法と比較法に関する実験結果について述べる．

5.1 実験データ

実データを用いた実験では，1994 年の 1 月から 6 月に記載された毎日新聞の国際面記事データ[3] を用いた．総文書数は 2695，語彙総数は 18070，トピック数は 45 個であった．記事にはあらかじめ人手によるトピックが付与されており，これを各記事の正解トピックと見なして各手法の評価を行う．記事データには各記事が掲載された日時，各記事の出現単語 ID とその出現頻度が記載されており，BOW 表現で記事がデータが記録されている．

ただし，この実データの文書に付与されたトピック情報は，必ずしも短期間の活性度に基づいて付与されたものではないため，ホットトピックの抽出を検証する上での適したコーパスとは言えないが，トピック抽出の性能面を検証する観点から実験を行った．

5.2 実 験 結 果

実データを対象とした提案法と比較法に関する実験結果を図 1 に示す．図 1 の横軸は，抽出されたトピック k のランキングに従い，横軸の左端を 1 位として K 位までのトピック k を評価の対象としていることを表している．また，縦軸は，各 K に対応し，抽出した記事群からトピックの正解データと最も良く一致した F 値を選び，正解トピック数 L で平均した平均 F 値 $F(K)$ を表したものである．また，最も性能が高い事例としては，正解トピック数 L が 45 であることから，抽出した記事群の数が 45 であるとき，全て正解トピックとして当てた場合である．したがって，抽出する記事群の数でもある K は，正解トピック数 45 以上の値を用い，正解トピック数の 10 倍までの記事群を抽出して実験を行った．

図 1 より，提案法は，他の手法と比較して高い性能を示すことがわかる．特に，他の手法と比べ，ランキング上位に相当する記事群と正解としたトピックが良く一致する傾向を示しており，提案法の高い優位性が示唆される結果となった．ただし，提案法の平均 F 値で最も高かった値は，52.7 に留まる結果となった．

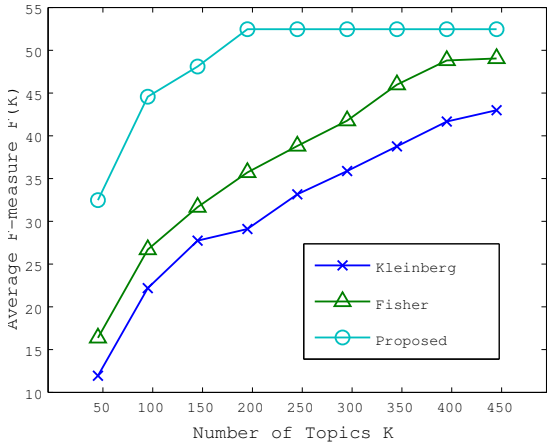


図 1 実データにおける SR 法と比較法との性能比較
Fig. 1 Comparison of the SR-method with comparative method for the real data.

また，表 2 に，提案法によりホットトピックとして見なされた上位 5 位までの記事群に含まれる，典型的な単語を列挙する．その記事群に典型的な単語群は，抽出した記事群のアノテーションとして，フィッシャー検定法により抽出したものである．

6. 人工データを用いた実験

本稿で利用した実データのコーパスは，必ずしもトピックの短期間活性度を考慮したものではなかったため，ホットトピックを抽出する性能を評価する上では適していなかった．

そこで，提案法の性能や特徴をより精細に検証するため，トピックが短期間に活性度が高くなる状況を疑似的に実現する人工データを作成し，実験を行った．人工データは，文書数

Vol.2009-MPS-76 No.51
Vol.2009-BIO-19 No.51
2009/12/18

表 2 ホットトピックのホットトピック期間及び関連単語群
Table 2 Hot topics's distance of hot topics and related word's group

コア番号	ホットトピック期間	抽出単語
1	6 月 1 日から 6 月 24 日	制裁, 北朝鮮, 朝鮮民主主義人民共和国, 核, カーター
2	5 月 3 日から 5 月 21 日	先行, ガザ, パレスチナ, 自治, 警察
3	2 月 6 日から 4 月 28 日	セルビア, サラエボ, 空爆, ボスニア, 勢力
4	6 月 17 日から 6 月 29 日	カーター, 日成, 南北, 洪, 板門店
5	4 月 5 日から 6 月 17 日	ゴラジュデ, 幸治, 上村, 幸彦, 町田

や単語数，文書出現期間などのパラメータを容易に変更できるという利点があり，データ構築を行う際にパラメータを細かく変更していくことによって，提案法や比較法の特徴をより明確に把握することができるものと思われる．

また，抽出された文書群に関する多重トピック性が各手法にどのような影響を及ぼすかを検証するため，人工データを作成する際，多重トピック性を導入した実験データを用意する．これらの観点を考慮した人工データの生成条件を以下に示す．

6.1 人工データの種類

トピックが短期間に活性化する状況，また多重トピック性を有する状況を人工的に実現するため，生成する人工データとして，まず,「単一トピックを持つトピック文書」と,「トピックを持たない一般文書」の二種類の文書を生成した．ここで，トピック文書が連続的に生成される期間を「トピック期間」と呼び，トピック文書が生成されない期間を「一般期間」と呼ぶ．これは，ある「トピック期間」と次の「トピック期間」の間に「一般期間」があるという意味である．ただし，多重トピックは,「単一トピックを持つトピック文書」の異なる種類どうしの生成期間を重ねることで疑似的に表現する．

また,「単一トピックを持つトピック文書」の異なる種類どうしの生成期間を重ねる際，その重複度の違いと重複させるトピックの種類の数に基づいて，人工データをそれぞれ四つの異なる状態として用意した．生成する文書数や総単語数，データ期間の長さは固定とし，トピック文書が生成される期間（トピック期間）を時間的に変動させる．本稿の実験においては,「単一トピックを持つトピック文書」の生成期間は，いずれの状態においても 5 日間とし,「トピックを持たない一般文書」の期間は常に 1 日間と固定した．

6.1.1 状態 1（トピックの重複期間なし）

状態 1 は，全期間において「単一トピックを持つトピック文書」の生成される「トピック期間」は，相互に独立しており，重複期間がない状態であるとする．つまり，トピック文書が生成される「トピック期間」と一般文書生成される「一般期間」が交互に繰り返されるが，相互の生成期間が重なることはない状態である．

6.1.2 状態 2（2 種類のトピック期間 50%重複）

状態 2 は，2 つのトピックが混在している重複状態が存在する場合である．ただし，状態 2 は，異なる 2 つのトピック期間 A と B が 50%重複している状態であり，その重複期間では，A と B という異なるトピックを持つトピック文書が同時に生成されることを意味し，A と B を合わせた期間が一つの「トピック期間」と見なされる．一方，トピック期間内で A と B が重複していない期間は，A と B がそれぞれ 1 種類のトピックの状態にあることを意味する．また，この場合，A と B の全体の期間が「トピック期間」となり，次の A' と B' の「トピック期間」の間に 1 日間の「一般期間」が存在する状態となる．

6.1.3 状態 3（2 種類のトピック期間 100%重複）

状態 3 は，2 つのトピック期間が 100%重複する状態とする．つまり，この場合の「トピック期間」では，常に二つのトピックが生成されている期間であることを意味する．また，この場合，重複期間全体が一つの「トピック期間」となり，次の「トピック期間」の間に 1 日間の「一般期間」が存在する状態となる．

6.1.4 状態 4（3 種類のトピック期間 100%重複）

状態 4 は 3 つのトピック期間が 100%重なっている状態であるとする．この場合の「トピック期間」では，トピック A を持つ文書，トピック B を持つ文書，トピック C を持つ文書が常に同時に生成されることを意味する．また，この場合も「トピック期間」どうしの間には 1 日間の「一般期間」が存在する．

6.2 人工データの生成法

四つの状態に基づく人工実験データは，トピックが事前に与えられている学習用データ D_s から事後分布 $P(\theta|D_s)$ が最大となる時のパラメータ値 θ を事前分布最大化学習によって各トピックごとに学習し，得られたパラメータ値の下で Naive Bayes モデルによって文書の生成を行う．ただし，Naive Bayes モデルによって文書を生成するには，生成するトピック文書ごとに，あらかじめパラメータを学習しておく必要があり，必然的に生成する人工データの基となる学習用データが必要となる．本稿における人工実験データに基づく実験では，学習用データとして 5 で使用した実データと同じ毎日新聞の国際面記事データを使用した．

まず，単一トピックを持つトピック文書の生成法について述べる．毎日新聞の国際面記事データには，45 個のトピックを持つ記事が存在しているが，その中からランダムに 24 個のトピックを選出し，それら 24 個のトピックを持つ文書から各トピックごとにパラメータの学習を行い，得られたパラメータの下でトピック文書の生成を行った．また，生成した人工

Vol.2009-MPS-76 No.51
Vol.2009-BIO-19 No.51
2009/12/18

データの全記事数，語彙総数は学習データに合わせ，それぞれ 2695，18070 とし，文書出現期間も 1994 年の 1 月から 6 月とした．

トピックを持たない一般文書の生成法としては，24 個のトピックから生成した文書を，それぞれの期間内で 50 記事ごと連続生成し，それらを混在させた文書群を学習用データとし，学習によって得られたパラメータを一般文書のパラメータと仮定し，その一般文書のパラメータを用いて特定のトピックを持たない疑似的な一般文書として生成を行った．

6.3 実験結果

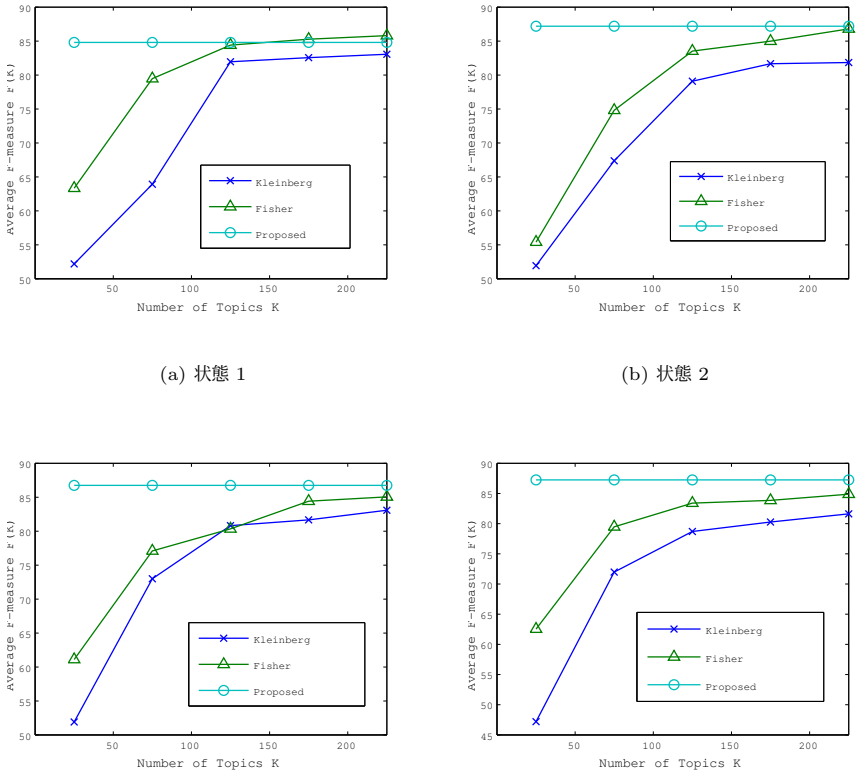
図 2 に人工データとして用意した四つの状態における提案法と比較法との性能比較を行った結果を示す．図 2(a)〜(d) のいずれも，横軸は，正解トピック数 L を 24 としたことから，K を 24 以上とし，10 倍となる 240 までの文書群を抽出する実験を行った．縦軸は，図 1 と同様，平均 F 値 $F(K)$ である．

まず，四つのいずれの状態においても，提案手法は比較法と比べて基本的に優位であることがわかる．図 2(a) の状態 1 で，K が 150 以上の部位では比較法が若干上回ることもあるが，抽出されたトピック k の上位では比較法よりも極めて高い優位性があることがわかる．

また，提案法は，状態 1 で平均 F 値が 84.8，状態 2 で 87.2，状態 3 で 86.7，状態 4 で 87.2 であることから，人工的なデータに対する結果ではあるものの，他の手法と比較して，短期間に活性化するホットトピックを高い精度で抽出できることもわかる．さらに，K の変化に関わらず，この平均 F 値は，変化していないことから，提案法の性能限界の範疇で，ランキングの上位に，正解トピックと最も良く一致する文書群を抽出できていることになるだけでなく，SR 法は，先に抽出したコア部がランキングの上位になる特徴があることから，抽出された文書群の早期の段階で，正解トピックと最も良く一致する文書群を抽出できたことにもなる．

また，トピックを短期間に独立させた状態 1 や，複数のトピックの重複性に変更を加えた状態 2 と 3，そして，重複するトピックの種類を増加させた状態 4 のいずれにおいても提案法は大きな性能の劣化がないことから，ランキングの上位で性能が大きく劣化する比較法と比べ，トピックが単独で存在する場合だけでなく，多重トピック性を持つ文書群についても，高いロバスト性をもつ可能性が示唆されているものと思われる．

以上の結果から，提案法は，短期間でトピックが活性化するホットトピックを高い精度で得る特徴があるだけでなく，ホットトピックを抽出する早期の段階で正解トピックと良く一致する文書群を抽出できるだけでなく，多重トピック性をもつ文書群に対してもロバストである点で，高い有効性があることが示唆された．これらの結果が明確にホットトピック性を有する文書群を持つ実データに対し，どのような結果をもたらすかについては，詳しい検証が必要となるが，人工データのパラメータを多角的に変更し，提案手法のより詳しい特徴を検証することも必要である．これらの検証については，今後の課題とする．



(a) 状態 1　(b) 状態 2

(c) 状態 3　(d) 状態 4

図 2 人工データにおける SR 法と比較法との性能比較
Fig. 2 Comparison of the SR-method with comparative method for the artificial data.

Vol.2009-MPS-76 No.51
Vol.2009-BIO-19 No.51
2009/12/18

7. おわりに

ネットワーク構造からコア部を抽出する手法としての SR 法を拡張し，文書ストリームデータから構築した文書ネットワークを対象として，ホットトピックを抽出する手法を提案した．また，実データと人工データによって提案法の評価を行った．新聞記事データを用いた実験においては，従来法と比較して，提案法が高い優位性があることを示した．また，人工データを用いた実験においては，提案法は，多項分布混合モデルにより時間領域局在性をもって生成される文書ストリームに対しては, 異なるホットトピック文書が時間領域において混在したとしても, 高精度にホットトピックを同定できることを示した.

参 考 文 献

1) Kleinberg. J.: Bursty and hierarchical structure in streams, *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-03)*, pp.91–101 (2002).
2) Saito, K., Ueda, N., Kimura, M., Kazama, K., and Sato, S.: Filtering search engine spam based on anomaly detection approach, *Proceedings of the KDD2005 Workshop on Data Mining Methods for Anomaly Detection*, pp.62–66 (2005).
3) 斉藤和巳, 木村昌弘, 上田修功: 文書トピックに関する認知科学的実験, 人工知能学会研究会資料 *(SIG-KBS-A405-10)* pp.57-62 (2005).
4) Zhao, Q., Mitra, P., and Chen, B.: Temporal and information flow based event detection from social text streams, *Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI-07)*, pp.1501–1506 (2007).
5) Swan, R., and Allan,J.: Automatic Generation of Overview Timelines, *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp.49–56 (2000).