The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

データ追加に基づく独立話題分析の話題変化検知手法の検討

Discussion on topic change detection method in Incremental Independent Topic Analysis

西垣 貴央 *1　　小野田 崇 *2
Takahiro Nishigaki　　Takashi Onoda

*1青山学院大学
Aoyama Gakuin University

We have proposed a topic extraction method that independent topic in increasing data. This algorithm extract independent topics form a small number of document data, update the independent topics when the new data comes. However, this method does not assume that the topic changes and the number of topics changes. In this paper, we propose a method to detect when the topic changes and when the number of topics changes. And we use some benchmark datasets to evaluate the proposed method. Evaluation results using benchmark data show that the proposed method is able to detect the topic changes and detect the number of topics changes.

1. はじめに

Web 上や個人所有のハードディスクドライブ (HDD) には大量の文書データが日々生成および蓄積されている．蓄積されている大量の文書データの中から，有益な知識を発見・抽出するためのテキストマイニング技術の一つである話題抽出について取り上げる．ここで話題とは， bag-of-words として与えられた大量の文書間で，複数の単語の共起によって表現される情報のことである [佐藤 15]．この話題を抽出する方法には様々な方法が存在するが，本稿では，独立性の高い話題を求める独立話題分析 [篠原 00] について考える．独立話題分析では，信号処理の分野で使用される独立成分分析 [Hyvärinen01] を用いて話題を求めている．ここで独立性が高い話題とは，話題間の相互情報量が小さい話題を示している．独立性が高い話題を求める利点として，より多くの情報量を持つ要約の作成が，容易にできる可能性が高いことが挙げられる．また，この独立話題分析を用いたシステムとして文書閲覧支援システム IT-DMS (Independent Topic-based Document Management System)[篠原 00] や，IT-DMS を改良した大量文書データに対する文書整理システム [田中 03] などがある．

逐次増加する文書データへ独立話題分析を適用できるように改良した，データ追加に基づく独立話題分析とよばれる方法 [Nishigaki17, 西垣 17] も提案されている．この方法では，初期データが抽出した独立話題を，新しい文書データが入ってくる度に更新することで，全てのデータを使用したときの話題と同等の話題を得ることが出来る．しかしこれら方法では，話題の中身が変化する場合や，話題の数が変わるということを想定していない．そこで本論文では，データ追加に基づく独立話題分析において話題の変化を検知する手法の検討を行う．

以下，2 章でデータ追加に基づく独立話題分析について簡単に紹介し，3 章で話題変化を検知する手法の検討を行う．4 章では，3 章で検討した方法をベンチマークデータに適用し，その評価を行う．最後に 5 章でまとめと今後の課題について述べる．

連絡先: 西垣 貴央，青山学院大学 理工学部 経営システム工学科，〒 252-5258 神奈川県相模原市中央区淵野辺 5-10-1 O-421c，nishigaki@ise.aoyama.ac.jp

2. Incremental Independent Topic Anlaysis

本章では，初期データのみで抽出した独立性の高い話題を，データが増加するたびに更新することで，全てのデータを用いて抽出した独立性の高い話題に近づける方法 [Nishigaki17, 西垣 17] について説明する．提案手法では，まず初期データに対して独立話題分析を適用し，独立性の高い話題を抽出する．次に，抽出した独立性の高い話題にもっとも影響を与えている文書データを抽出した独立性の高い話題の数だけ抜き出し，抽出した独立性の高い話題と抜き出した文書データ以外のデータ全てを削除する．それから，データが追加されると追加されたデータと抽出した独立性の高い話題と抜き出した文書データを合わせ一つのデータとする．その合わせたデータを用いて，抽出した独立性の高い話題を FPICA を用いて更新する．再び，更新した独立性の高い話題と抜き出した文書データのみを残して他のデータ全てを削除する．これを繰り返すことで，最終的に独立性の高い話題を得る．提案したデータ追加に基づく独立話題分析のアルゴリズムを以下に述べる．共通の変数として，話題インデックスを $t \in \{1, \ldots, k\}$，文書インデックスを $d \in \{1, \ldots, n\}$，単語インデックスを $w \in \{1, \ldots, m\}$ とする．

1. 初期データの各文書中の各単語の頻度の行列 $\mathbf{A}$ を作成し，単語数の偏りが出ないように正規化を行い $\tilde{\mathbf{A}}$ を得る．
2. $\tilde{\mathbf{A}}$ に対して特異値分解を行い，$\tilde{\mathbf{A}}$ を次のように分解する $\hat{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{A}}\hat{\mathbf{V}} = \hat{\mathbf{S}}$．ここで，$\hat{\mathbf{S}}$ は特異値の対角行列である．
3. ステップ 2. で得た行列 $\hat{\mathbf{U}}$ と $\hat{\mathbf{S}}$，$\hat{\mathbf{V}}$ を，$\hat{\mathbf{S}}$ の値の大きい順に k 個の成分を抜き出し，行列 $\mathbf{U}$，$\mathbf{S}$，$\mathbf{V}$ を作成する．
4. k 次元空間での話題を示す $k \times m$ の行列 $\mathbf{X}$ を次の式で定義する $\mathbf{X} = \mathbf{S}^{-1/2}\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{A}}$.
5. 各話題の独立性最大化：ステップ 4. で得られた話題に対して，FPICA[Hyvärinen99] に基づいて最大の独立性を与えるための回転行列 $\mathbf{R}$ を次のように決定する．
   (a) $\mathbf{R}$ の初期値を $k \times k$ の零行列とする $\mathbf{R} = \mathbf{0}$.
   (b) 単位行列 $\mathbf{I} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k)$ の $t \in \{1, \ldots, k\}$ 列目の列ベクトルを $\mathbf{e}_t$ として，回転行列 $\mathbf{R}$ の t 列目

$\mathbf{r}_t = (r_{1,t}, r_{2.t}, \ldots r_{k,t})^{\mathrm{T}}$ に代入する $\mathbf{r}_t = \mathbf{e}_t$．ここで，$\mathbf{e}_1 = (1, 0, \ldots, 0)^{\mathrm{T}}$，$\mathbf{e}_2 = (0, 1, 0, \ldots, 0)^{\mathrm{T}}$ の $k \times 1$ の単位ベクトルである．

(c) $\mathbf{r}^{(old)}$ に $k \times 1$ の零ベクトルを代入して，$\mathbf{r}^{(old)}$ を次のように初期化する $\mathbf{r}^{(old)} = (0, 0, \ldots, 0)^{\mathrm{T}}$．

(d) $\mathbf{r}_t$ を次の式で更新する $\mathbf{r}^{(old)} = \mathbf{r}_t$，$\mathbf{r}_t = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{r}_t).^3 - 3\mathbf{r}_t$．$(\mathbf{X}^{\mathrm{T}}\mathbf{r}_t).^3$ は $\mathbf{X}^{\mathrm{T}}\mathbf{r}_t$ の行列要素の 3 乗を表す．

(e) $\mathbf{r}_t$ を次の回転行列化を行う $\mathbf{r}_t = \mathbf{r}_t - \mathbf{R}\mathbf{R}^{\mathrm{T}}\mathbf{r}_t$，$\mathbf{r}_t = \mathbf{r}_t / \|\mathbf{r}_t\|$．

(f) $\|\mathbf{r}_t \pm \mathbf{r}^{(old)}\|$ が閾値以上ならば，ステップ 5d. へ．閾値より小さければステップ 5g. へ．

(g) $t < k$ ならば，t を 1 つ増やして，ステップ 5b. へ．$t = k$ ならば，その時の $\mathbf{R}$ を回転行列として，ステップ 6. へ．

6. 独立な話題中での単語の重要度 ${}^*\mathbf{V}$ と独立な話題中の文書の重要度 ${}^*\mathbf{U}$ を下記により計算する．

$$ {}^*\mathbf{V} = \mathbf{V}\mathbf{R}, \quad {}^*\mathbf{U} = \mathbf{U}\mathbf{R} $$

7. 得られた各話題に対して $\mathbf{u}_k$ の絶対値が最も大きい文書データを抜き出す．

8. ステップ (6) で得られた話題とステップ (7) のデータを除いて他の全てのデータを削除する．

9. 新しく追加されたデータに，ステップ (8) で残ったデータを連結する．

   - このとき新しく追加されるデータは，ステップ (1) での初期データに対して非常に小さいものとする．

10. データが増加した後の新たな独立な話題を求める回転行列を $\tilde{\mathbf{R}}$ とし，その初期値をステップ (6) で得られた $\mathbf{R}$ とする．

11. ステップ (9) のデータに対して，ステップ (10) の回転行列 $\tilde{\mathbf{R}}$ を独立話分析のステップ (5) と同様に FPICA に基づいて新しい回転行列を得る．

12. データが加えられた後の，新しい独立な話題が得られる．

このステップ (7) からステップ (12) をデータが増加する度に繰り返すことで，データ追加に基づいた独立性の高い話題を抽出することができる．

データ追加に基づく独立話題分析のイメージを図 1 に示す．この方法によって，増加するデータに対しても全てのデータを同時に用いることなく，独立な話題を求めることができる．

3. 話題変化を検知する手法の検討

独立話題分析では，抽出された話題中での単語の重要度 ${}^*\mathbf{V}$ と話題中の文書の重要度 ${}^*\mathbf{U}$ によって，話題が表現される．抽出された話題の $\mathbf{v}_t$ と $\mathbf{u}_t$ はそれぞれ，t 番目の話題の単語の重要度と文書の重要度を表すベクトルである．また，τ 回目に得られた t 番目の話題の単語重要度を $\mathbf{v}_t^{(\tau)}$ と表現することにする．$\mathbf{u}_t^{(\tau)}$ についても同様である．

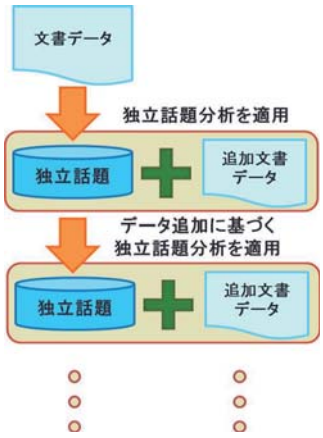

図 1: データ追加に基づく独立話題分析のイメージ

本稿では，1) τ 回目に抽出された話題と $\tau + 1$ 回目に抽出された話題が，同じ話題なのか異なる話題なのかを検知することと，2) τ 回目までは存在しなかった話題が $\tau + 1$ 回目から新たに現れたことを検知することを，目標とする．

3.1 話題の変化の検知

抽出された話題の $\mathbf{v}_t^{(\tau)}$ および $\mathbf{u}_t^{(\tau)}$ と，$\mathbf{v}_t^{(\tau+1)}$ および $\mathbf{u}_t^{(\tau+1)}$ を比較することで，話題の強弱を調べることができると考えた．比較には，各話題はすべてベクトルで表現されるので，コサイン類似度を用いて以下の式で表す．

$$ \cos(\mathbf{v}_t^{(\tau)}, \mathbf{v}_t^{(\tau+1)}) = \frac{\mathbf{v}_t^{(\tau)} \cdot \mathbf{v}_t^{(\tau+1)}}{|\mathbf{v}_t^{(\tau)}||\mathbf{v}_t^{(\tau+1)}|}, \quad t \in \{1, \cdots, k\} $$

k 個の話題それぞれで，$\mathbf{v}_t^{(\tau)}$ と $\mathbf{v}_t^{(\tau+1)}$ のコサイン類似度を計算することで，その話題を構成する単語の重みの違いが明確になる．コサイン類似度の値が 1 もしくは -1 の場合は，全く同じ強さを持った話題となり変化はなかったことを示す．一方で，コサイン類似度の値が 0 に近づくにつれて，異なる話題となり変化があったことを示す．この値の絶対値がある値 α よりも小さければ，τ 回目の話題 t と $\tau + 1$ 回目の話題 t は異なる話題であるということが言える．また，抽出された話題が変化がなかったとしても，その話題に分類される文書数の変化によって，その話題の強弱を調べることができる．話題の強弱とは，その話題が強い場合，多くの文書がその話題に分類されていることを指し，反対に弱い場合は，その話題に分類される文書数が少ないことを指す．話題への文書データの分類は，次の式を用いて行う．式中の $\mathbf{a}_i$ は i 番目の文書を表しており，$\mathbf{u}_i^{\mathrm{T}(\tau)}$ は，τ 回目における i 番目の文書の話題への影響度を表したベクトルとなっている．つまり，$\mathbf{u}_i^{\mathrm{T}(\tau)}$ の値が最大値を取るのが t 番目の話題のとき，その文書 $\mathbf{a}_i$ は話題 k に分類される．$\mathcal{T}_t$ は話題 t に分類された文書数を表す．

$$ \mathcal{T}_t = \{\mathbf{a}_i | \arg\max_i \mathbf{u}_i^{\mathrm{T}(\tau)} = t\}, \, i \in \{1, \ldots, \text{追加される文書数}\} $$

これをすべての話題で行い，各話題へ分類されている文書数の多さで話題への強弱を推定する．

3.2 新規話題の検知方法

新しい話題の出現の検知方法にも，コサイン類似度を使用する．$\tau + 1$ 回目で抽出されたある話題が，τ 回目の全ての話題と一定以上離れている場合，$\tau + 1$ 回目で抽出されたある話

表 1: LA Times に話題数 6 で独立話題分析を行った時の各話題に分類された文書数

話題番号	話題 1	話題 2	話題 3
文書数	898	704	853
話題番号	話題 4	話題 5	話題 6
文書数	432	924	2468

題は新規話題であると推測できる．

$$\cos(\mathbf{v}_t^{(\tau)}, \mathbf{v}_j^{(\tau+1)}) = \frac{\mathbf{v}_t^{(\tau)} \cdot \mathbf{v}_j^{(\tau+1)}}{|\mathbf{v}_t^{(\tau)}||\mathbf{v}_j^{(\tau+1)}|}, \quad (t, j \in \{1, \cdots, k\},\ t \neq j)$$

このコサイン類似度を計算し，この値の絶対値がある値 α よりも小さければ，その話題 j は新規話題であるということが言える．

4. 実験

4.1 実験に使用するデータについて

実験でベンチマークデータとして Los Angeles Times (LA Times) の新聞データ [Zhong03] で文書数は 6279，単語数は 31472，話題数は 6 の文書データを使用する．

その新しい文書の中に新規話題がある場合を考える．しかし，LA Times のデータは時系列データではないため，話題の変化および話題数の変化については情報がない．そこで本実験ではデータの準備として，最初に LA Times に話題数 6 で独立話題分析を行い，各文書データを各話題に分類した．分類には抽出した独立性の高い話題とのコサイン距離を用いた．その時，各話題での文書数は，表 1 となった．表 1 の数で各文書が分類され，話題 5 を除いた話題 1，話題 2，話題 3，話題 4，話題 6 の 5 個の話題に属する文書数 5355 と話題 5 の文書数 924 を別に扱う．

本実験では，データ追加に基づく独立話題分析では，初期データとして全データの 50% を用いて初期の独立話題を抽出する．その後，必要最低限の情報だけを残して，新しい文書数 200 を逐次追加していく．初期データとして全データの 50%（3139 文書）を話題 5 を除いた文書からランダムに選択し，初期の独立話題を抽出を行った．その後，200 ずつデータを追加していくため，全部で 16 回データの追加が行われた．追加されるデータの 4 回目までは，既存の話題のみが含まれるデータとして，5 回目から少しずつ新規話題（話題 5）が含まれるデータを追加することとした．

以上のデータを用いて，提案する方法で話題の変化および話題数の変化を検知することができるか検証を行う．

4.2 結果と考察

まず話題の変化の検知について結果を示す．話題の変化の検知では，同じ話題と考えられる話題同士のコサイン類似度によって決定する．表に，τ 回目と $\tau+1$ 回目での話題でのコサイン類似度を角度に変換した値を示す

表 2 では $\tau = 0$ で非常に大きな角度になっていることが分かる．これは，使用しているデータ数が初期データでは 3139 あるのに対して，追加以降は 212 になるためであると考えられる．そのため，1 回目は除外する．$\tau = 4$ 回目までは既存の話題のみを追加しているため，5 回目と 6 回目の角度に注目すると，$\tau = 5$ 回目では話題 4 と話題 6 が，6 回目では話題 5 と話題 6 のときに，10 度以上離れていることが分かる．この時，

表 2: τ 回目と $\tau+1$ 回目での話題の角度

τ	話題 1	話題 2	話題 3	話題 4	話題 5	話題 6
0	16.41	3.39	4.35	4.14	12.59	18.48
1	7.81	4.05	8.00	0.14	0.25	0.19
2	8.12	2.28	7.53	5.45	4.01	1.43
3	4.88	6.55	4.95	3.8117	4.74	0.26
4	4.96	3.85	3.98	11.38	4.62	13.24
5	3.98	3.2383	4	0.94738	11.71	11.40
6	0.02	6.24	4.19	3.89	1.946	4.50
7	5.04	5.66	2.33	4.32	1.17	2.06
8	3.8	3.84	4.97	2.09	0.22	0.137
9	4.99	3.84	3.37	3.93	6.38	1.94
10	0.16	5.23	4	2.78	1.81	4.075
11	4.35	3.31	1.81	2.13	0.42	1.21
12	4.18	4.95	3.77	0.37	1.93	1.77
13	3.77	8.03	4.41	2.2	8.08	0.31
14	4.16	8.07	3.76	3.48	8.15	3.75
15	3.76	5.26	3.99	3.76	2.15	3.47

話題が変化したのではないかと考えられる．以上のことより，α を 0.9848 とすることで，10 度以上離れたら話題が変化したとすることができる．

次に話題の強弱について，各話題でどの程度文書が含まれているかを表 3 を用いて考察する．

表 3 をみると，各話題に含まれる文書数は各回ごとにかなり異なることが分かる．話題が変化したと考えられる，5 回目の話題 4 と話題 6，6 回目の話題 5 と話題 6 を見ると，大きく数が増減していることが分かる．各話題に分類される文書数が大きく変わったから，話題の変化が起こった様にも考えられるが，2 回目の話題 6 と 3 回目の話題 6 に分類される文書数を見ると，ここも大きく減少している．しかし，表 2 の $\tau = 2$ の話題 6 を見ると，1.43 と非常に小さいことが分かる．これは話題数は大きく減っているが，話題の内容は変化していないことを示している．

最後に，提案した新規話題の検知方法について検証を行う．本実験では，$\tau = 4$ のときに新しい話題が出現している．その時の各話題での角度を表 4 に示す．表 4 では，各話題同士の角度を表しており，0 度もしくは 180 度のときは同じ話題であるということを示している．90 度のときは，独立した話題であると言える．全く同じ話題でコサイン類似度を計算した場合，ある 1 個の話題が 0 度もしくは 180 度となり，他全ての話題が 90 度となる．合計の行では，それぞれの話題が 90 度あるいは 0 度もしくは 180 度からどの程度離れているのかの合計値を示している．そこで話題が大きく変化して新規話題が出現した場合，この合計値の値が大きな値となると考えられる．表 4 の合計値を確認すると，$\tau+1$ の話題 6 は τ の全ての話題から 38 度以上離れていること示しており，ここで $\tau+1$ 回目の話題 6 が新規話題である可能性が高いことを示している．

以上の結果より

5. おわりに

本論文では，データが逐次増加する場合における独立話題分析において，データが増加した時に話題の変化および話題数の変化を検知する手法を提案した．また，LA Times のデー

表 3: τ 回目での各話題に分類された文書数

τ	話題 1	話題 2	話題 3	話題 4	話題 5	話題 6
1	6	4	48	2	67	85
2	19	15	34	22	12	110
3	72	55	30	12	41	2
4	14	120	10	46	17	5
5	32	28	24	3	37	88
6	34	71	26	73	4	4
7	41	74	39	4	5	49
8	35	11	43	8	98	17
9	109	42	11	15	12	23
10	22	47	27	82	25	9
11	39	68	41	22	5	37
12	5	29	84	52	4	38
13	28	18	65	42	49	10
14	7	20	94	53	29	9
15	7	6	18	102	40	39
16	7	17	65	43	16	4

表 4: $\tau = 4$ の時と $\tau = 5$ の時の各話題との角度

話題	$1(\tau+1)$	2	3	4	5	6
$1(\tau)$	93.43	87.85	176.02	89.93	90.12	90.25
$2(\tau)$	93.11	86.79	89.53	80.15	93.73	13.24
$3(\tau)$	175.04	90.47	86.85	89.79	90.08	94.35
$4(\tau)$	89.95	3.85	87.78	90.75	89.66	92.82
$5(\tau)$	90.41	90.09	90.13	90.54	175.38	94.79
$6(\tau)$	90.96	90.31	90.53	11.38	91.51	103.02
合計	12.93	10.04	10.47	22.80	10.40	38.79

タを使用して，提案手法によって話題の変化および話題数の変化の検知が行うことができるのか検証した．検証の結果，提案した話題の変化の検知および新規話題の検知方法の両方共 LA Times の場合はうまく検知できていると考えられる．しかし，LA Times は時系列データではないため，理想環境としてデータの追加を行っており，時系列情報が含まれるデータや他のベンチマークデータでも良い結果が得られるとは限らない．また，話題の変化や新規話題の検知で用いる角度は，どの程度を離れていると変化が起こった，あるいは新規話題であるということができるのか，LA Times の結果だけではわからないため，他のデータに対して行う必要がある．さらに，本稿ではコサイン類似度の値のみでの分析しか行っていないため，それが本当に新規話題を示しているのか，たまたま異なる話題が得られてしまったのか検証が行えていないため，話題に含まれる単語を比較して分析を行う必要がある．

謝辞

本研究の一部は，青山学院大学アーリーイーグル研究支援制度の支援によって行われた．

参考文献

[Hyvärinen99] Aapo Hyvärinen: Fast and Robust Fixed-Point Algorithms for Independent Component Analysis, IEEE Trans. on Neural Networks, Vol.10, No.3 (1999).

[Hyvärinen01] Aapo Hyvärinen and Juha Karhunen and Erkki Oja: Independent Component Analysis, John Wiley & Sons (2001).

[Nishigaki17] Takahiro Nishigaki, Katsumi Nitta, Takahis Onoda: Incremental Learning of Independent Topic Anlysis, International Journal of Computer, Electrical, Automation, Control and Information Engineering, Vol. 11, No. 2, pp.191–197 (2017).

[Lichman13] M. Lichman: UCI machine learning repository, http://archive.ics.uci.edu/ml , 2013.

[Zhong03] Shi Zhong and Joydeep Ghosh: A comparative study of generative models for document clustering, Data Mining Workshop on Clustering High Dimensional Data and Its Applications (2003).

[佐藤 15] 佐藤 一誠: トピックモデルによる統計的潜在意味分析, 自然言語処理, 第 8 巻, コロナ社 (2015).

[篠原 00] 篠原 靖志: 文書データベースの主要話題の発見と変化の追跡を行う文書閲覧支援システムの開発, 電力中央研究所報, R99036 (2000).

[田中 03] 田中 真人, 篠原 靖志: 重要話題発見のための大量文書自動整理システム, 電力中央研究所報告, R02015 (2003).

[西垣 17] 西垣 貴央, 新田 克己, 小野田 崇: 逐次データ追加がある状況下での独立話題分析とその実験的特性分析, 人工知能学会全国大会 (第 31 回) 論文集, 1E3-1, (2017).