1E3-1

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

Experimental characteristic analysis of Independent Topic Analysis when the number of data increases

西垣 貴央 *1　新田 克己 *1　小野田 崇 *2
Takahiro Nishigaki　Katsumi Nitta　Takashi Onoda

*1東京工業大学
Tokyo Institute of Technology

*2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. In this paper, we use some benchmark datasets to evaluate the proposed method, and we show the feature of the proposed method. Evaluation results using benchmark data show that the proposed method is able to extract the close topic as the topic extracted using all the data, and difficult to extract the same topic when the number of topics is small and the number of data included in each topic varies greatly.

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．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, 村田 05] 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．

しかし，この独立話題分析では逐次増加する文書データへの適用は難しい．なぜなら，独立話題分析では独立な話題を求めるために，全ての文書データを使用して話題を求めるため，新しい文書データが入ってから再び独立な話題を求める場合，再び全ての文書データを使用する必要があるためである．この課題に対して，解決を試みた方法としてデータ追加に基づく独立話題分析 [西垣 16, Nishigaki17] を提案した．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．

本論文では，提案したデータ追加に基づく独立話題分析の有効性を，複数のベンチマークデータに適用し，その結果の考察を行い，提案手法の特徴を報告する．

以下，2 章で独立話題分析について簡単に紹介し，3 章では提案したデータ追加に基づく独立話題分析について述べる．4 章では，提案したデータ追加に基づく独立話題分析をベンチマークデータに適用し性能評価を行う．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: 連絡先: 西垣貴央, 東京工業大学大学院 総合理工学研究科 知能システム科学専攻, 〒 226-8502 神奈川県横浜市緑区長津田町 4259 J2-53, nishigaki@ntt.dis.titech.ac.jp

2. 独立話題分析

本章では，篠原によって提案された独立話題分析 [篠原 00] について簡単に紹介する．以下，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．

まず独立話題分析における諸概念を簡単に述べる．$\mathbf{V}$ は $m \times k$ の行列であり，“単語 w の話題 t での重要度” を示す．$\mathbf{U}$ は $n \times k$ の行列であり，“文書 d の話題 t での重要度” を示す．同様に $\mathbf{A}$ は $n \times m$ の行列であり，“文書 d 中での単語 w の頻度” を示す．また，ここで話題間の独立性を評価する指標として代表的な指標の一つである高次統計量の尖度（同一の平均・分散を持つ正規分布との 4 次モーメントの差）を使用する．尖度を使用した “話題の単語集中度” は次のように定義する．

$$\sum_{w}^{m} v_{w,t}^4 P(w) - 3\left(\sum_{w}^{m} v_{w,t}^2 P(w)\right)^2$$

$v_{w,t}$ は行列 $\mathbf{V}$ の w 行 t 列の要素である．ここで $P(w)$ は単語 w の全文書中での出現確率を示す．話題の単語集中度の値が大きいということは，大半の単語や文書の重要度が 0 の近くにあり，重要度の大きい単語や文書が少数しかないことを示す．すなわち，少数の単語や文書のみでその話題を表すことができる．話題間の独立性の強さは，各話題における集中度の二乗和によって定義する．この値が大きい場合，各話題に重要度の大きい単語や文書が集中していることを示すので，話題間の独立性は高くなる．

独立話題分析は，これらの諸概念を用いて文書データから，話題の単語集中度が最大となる $\mathbf{V}$ を求めるものである．あらかじめ求めたい話題数 k が与えられており，文書 d 中での単語 w の頻度を示す行列 $\mathbf{A}$ から，各話題の重要度 $\mathbf{V}$ や $\mathbf{U}$ を座標軸とする k 次元空間を求め，その空間に各文書と各単語を配置する．このとき，各話題は正規直交性を満たしている．独立話題分析では，文書に対する点の近くに，その文書中に現れる単語に対応する点もあるという最適な配置を実現する．そして最適な配置の中で，各話題の独立性が最大となる配置 $^*\mathbf{V}$ を回転行列 $\mathbf{R}$ を用いて求める．

次にそのアルゴリズムを示す．

1. 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. 独立な話題中での独立の重要度 ${}^*\mathbf{V}$ 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．

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

以上の独立話題分析によって得られる話題は，図 1 のようになり独立性の高い話題を得ることができる．例として，Los Angeles Times (LA Times) の論文データ [Karypis02, Zhong03] に対して話題数 6 で独立話題分析を行うと表 1 のように話題が抽出できる．表 1 で示す話題を構成する重要単語とは，単語の話題での重要度を示す行列 $\mathbf{V}$ の各話題（各列）において要素の絶対値が大きいもの 5 個を示している．

この独立話題分析は，上記のように与えられた文書データ全てを用いて，独立性の高い話題を得る方法である．そのため独立話題分析は，数が増加していくデータに独立話題分析を適用するのは困難である．なぜなら，独立話題分析を逐次的に増加するデータに適用する場合，データが増加する度に全てのデータを使用しなくてはならないからである．そこで，データが逐次増加する場合においても，独立性の高い話題を求めることが出来る方法が提案されている．その方法について次章で述べる．

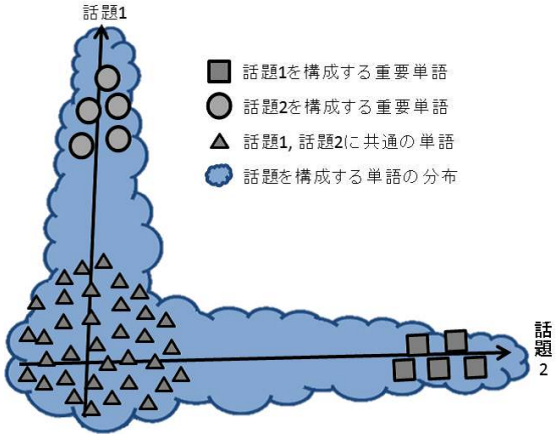

図 1: 独立話題分析のイメージ

表 1: LA Times に独立話題分析を適用して得られた 6 個の話題を構成する重要単語

話題	重要度が高い単語				
	$w=1$	$w=2$	$w=3$	$w=4$	$w=5$
1	scor	game	lead	quarter	rebound
2	soviet	afghanistan	israel	foreign	militari
3	aleen	macmin	art	entertain	report
4	bush	polic	budget	senat	towert
5	million	earn	bank	quarter	billion
6	polic	counti	officr	orang	citi

3. データが逐次増加する場合の独立話題分析

本章では，我々が提案した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 [西垣 16, Nishigaki17] 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．

1. 初期データに対して独立話題分析を行い任意の数 k 個の独立性の高い話題を抽出する．

2. 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．

3. ステップ (1) とステップ (2) のデータを除いて他の全てのデータを削除する．

4. 新しく追加されたデータに，ステップ (3) で残ったデータを連結する．

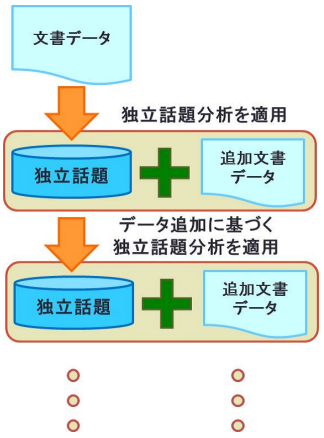

図 2: データ追加に基づく独立話題分析のイメージ

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

5. 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 (1) で得られた $\mathbf{R}$ とする．

6. ステップ (4) のデータに対して，ステップ (5) 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．

7. 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 (2) からステップ (7) 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 2 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．

4. 実験

本章では，提案されたデータ追加に基づく独立話題分析を複数のベンチマークデータに適用し得られた話題の特性および，提案手法の特性を検証するための実験を行う．

4.1 実験に使用するデータおよび評価方法

実験には，話題数が多いデータの Los Angeles Times (LA Times) の新聞データ [Karypis02, Zhong03] で文書数は 6279，単語数は 31472，話題数は 6 の文書データと，話題数が少ないデータの KOS Blog のブログデータ [Lichman13] で文書数は 3430，単語数は 6906，話題数は 2 の文書データを用いる．

実験では，これらのベンチマークデータからランダムに 50% のデータ抽出しをそれを初期データとした．また，追加データ初期データより十分小さい値を想定しているため，データ数は 100 とした．実験の初期データ数は LA Times で文書数 3139，KOS Blog の文書数は 1715 となる．

評価には，ベンチマークデータに提案手法を適用して得られた話題を構成する重要単語と全てのデータを使用して得られる話題を構成する重要単語の重複率の比較を行う．またそのときの重要単語の上位 5 個の比較も行う．話題を構成する重要単語とは，単語の話題での重要度を示す行列 $\mathbf{V}$ の各話題（各列）において要素の絶対値が大きいものを示している．重要単語の重複率とは，提案手法によって得られた話題を構成する重要単語と全てのデータを使用して得られた話題を構成する重要単語がどの程度一致しているかの値となる．1 が最大で 1 のとき，二つの話題で使用されている重要単語は完全に一致していることを指す．

表 2: ベンチマークデータの 50% を初期データとして，提案手法を適用し得られた話題を構成する重要単語と全てのデータに独立話題分析を適用して得られた話題を構成する重要単語の重複率

重要単語数の割合	1%	10%	20%
LA Times（初期データ時）	0.639	0.636	0.668
LA Times（最終時）	0.642	0.639	0.683
KOS Blog（初期データ時）	0.913	0.778	0.705
KOS Blog（最終時）	0.648	0.598	0.601

表 3: LA Times の 90% を初期データとし，これに提案手法を適用して得られた話題を構成する重要単語と全てのデータに独立話題分析を適用して得られた話題を構成する重要単語の重複率

重要単語数の割合	1%	10%	20%
LA Times（初期データ時）	0.836	0.871	0.886
LA Times（最終時）	0.897	0.900	0.910

4.2 結果と考察

前節で述べた実験設定で行った実験結果を示す．まず，ベンチマークデータに提案手法を適用して得られた話題を構成する重要単語と全てのデータに独立話題分析を適用して得られた話題を構成する重要単語の重複率を表 2 に示す．

表 2 の上段を見ると，LA Times の場合，重要単語数の割合の上位 1%，10%，20% のいずれの場合も，最終的に得られる話題の方が，初期データ時と比較して値が高いため，提案手法はデータが追加されると，全てのデータに独立話題分析を適用して得られる話題に近いていることがわかる．しかし値の上昇幅は非常に小さい．値の上昇幅が非常に小さい理由として考えられることは，LA Times のデータに含まれる各話題に属する文書数が均等ではないためだと考えられる．LA Times で最も文書数が多い話題と最も文書数が少ない話題とでは，4 倍程度その数が異なる．そのため話題数が 6 個と多く，かつ初期データ数が 50% 程度と少ない場合，選択した初期データによって最初に抽出される話題に大きく偏りが出てしまう．そこで，LA Times の初期データの数を 90% とした場合の単語の重複率を表 3 に示す．表 3 を見ると，初期データの数が多くなることで初期データ内に含まれる話題に偏りが少ない場合は，重複率の値も非常に高くなり，かつデータを追加していくことで重複率の上昇率も表 2 のときと比較すると上がっていることがわかる．以上のことから，提案するデータ追加に基づく独立話題分析を適用することで，全てのデータから求めた独立話題に近づけるためには，初期データ内に求めたい全ての話題の文書が含まれている必要があることがわかる．

次に KOS Blog の場合について考察する．表 2 の下段を見ると，KOS Blog の場合，重要単語数の割合の上位 1%，10%，20% のいずれの場合も，初期データ時で得られる話題の方が，データが追加されて最終的に得られる話題のときよりも値が高いことがわかる．この理由として考えられることは KOS Blog

表 4: KOS Blog の 50% を使用して，データ追加に基づく独立話題分析を話題数 2 で適用して得られた話題を構成する重要単語上位 5 個

重要度が高い上位単語	各話題の重要度が高い単語	
	1	2
$w=1$	bush	november
$w=2$	iraq	poll
$w=3$	war	account
$w=4$	president	electoral
$w=5$	kerry	governor

表 5: KOS Blog の全てのデータに独立話題分析を適用して話題数 2 で適用して得られた話題を構成する重要単語上位 5 個

重要度が高い上位単語	各話題の重要度が高い単語	
	1	2
$w=1$	bush	november
$w=2$	iraq	house
$w=3$	war	senate
$w=4$	president	electoral
$w=5$	administration	account

は政治に関する話題に限定されたデータであるため，使用されている単語数が LA Times と比較すると少ない．そのため，少ない初期データで求めた話題が全てのデータを用いて得た話題と十分近い話題が得られてしまっていることが考えられる．初期データで求めた話題が，全てのデータを使用して求めた話題と十分近い話題が得られていることは，表 4 および表 5 を見るとわかる．表 4 は 50% の初期データに提案手法を適用して得られた話題を構成する重要単語の上位 5 個を示し，表 5 は全てのデータを独立話題分析に適用して得られた話題を構成する重要単語の上位 5 個である．これを見ると二つの表で示した単語はほとんど同じであることがわかる．提案手法では，初期データで求めた話題は，新たに追加されたデータに基づいて更新していくため，新たに追加されたデータの影響を大きく受けてしまう．そのため，初期データで求めたときの値が最も大きな値を示してしまうと考えられる．

以上の二つのベンチマークデータへ，データ追加に基づく独立話題分析を適用した．実験では初期データだけを使用して得られた話題と，データが追加されて最終的に得られた話題と全てのデータを使用して得られた話題とでそれらを構成する重要単語の比較を行った．その結果より，提案手法によって得られた話題は，データが追加される度に更新することで，全てのデータを用いて得られた独立性の高い話題に近づいていくことが示せた．しかし，提案手法を適用するデータは話題数は多い方が望ましく，かつ初期データに使用するデータ内に，求めたい全ての話題に関するデータが含まれている必要があることがわかった．

5. おわりに

本論文では，提案されたデータが逐次増加する場合における独立性の高い話題を求める方法をいくつかのベンチマークデータに適用して得られる話題の特性について検証した．また，提案手法を適用することが向いているデータ，向いていないデータなどの，提案手法の特性についての考察も行った．その結果，提案手法はデータを追加することで全てのデータを使用して得られる独立性の高い話題に近づけることができることが示せた．しかし，提案手法を適用するデータは話題数が多いほうが望ましく，かつ初期データに使用するデータ内に，求めたい全ての話題に関するデータが含まれている必要があるということがわかった．

今後の課題として，提案手法は，データが増加したときに新しい話題が増えるという場合に対処できないため，データが増加したときに新たな話題が増加すればそれを検知し，求める話題数が変更できる方法について検討する必要がある．

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).

[Karypis02] George Karypis: CLUTO - A Clustering Toolkit, http://glaros.dtc.umn.edu/gkhome/views/cluto/ , Department of Computer Science and Engineering, University of Minnesota (2002).

[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).

[村田 05] 村田 昇: 入門 独立成分分析, 東京電機大学出版 (2005).

[西垣 16] 西垣 貴央, 新田 克己, 小野田 崇: データ追加に基づく独立話題分析の提案, 人工知能学会全国大会 (第 30 回) 論文集, 2J3-5, (2016).