2J3-5

データ追加に基づく独立話題分析の提案

Incremental Independent Topic Analysis

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電力中央研究所 Central Research Institute of Electric Power Industry

The number of document data has been increasing since the spread of Internet. Independent Topic Analysis (ITA) was proposed as one method to analyze the document data. ITA is a method for extracting the independent topics from a large number of document data. However, applying ITA to the increasing number of document data is a hard because temporal and spatial cost is large. In this paper, we present Incremental ITA (IITA) which estimates the independent topics from increasing number of document data. IITA is a method for extracting the independent topics from a small number of document data. In addition, IITA update the independent topics when the document data is added. And we show the result that applied IITA to benchmark datasets.

1. はじめに

近年，ノート PC やスマートフォンなど安価で高性能なデバイスの普及や，インターネット利用の一般化に伴い，Web ページや電子ニュース，またブログやソーシャルネットワークサービス (SNS) などが広く活用されている．そのため，Web 上や個人のハードディスクドライブ (HDD) には大量の文書データが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として，確率的生成モデルに着目して話題を抽出する方法と，抽出する話題間の関係に着目して話題を抽出する方法が存在する．確率的生成モデルに着目して話題を抽出する方法として，Hofmann の提案した PLSA (Probabilistic Latent Semantic Analysis)[Hofmann 99] や Blei らの提案した LDA (Latent Dirichlet Allocation)[Blei 03] などが研究されている．一方で話題間の関係に着目して話題を抽出する方法としては，LSI (Latent Semantic Indexing)[Deerwester 90] や独立話題分析 [篠原 99] がある．

本稿では，独立性の高い話題を求める独立話題分析 [篠原 99] 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 [Hyvärinen 01, 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．一方で，膨大な量の文書データに対して独立話題分析を適用することは，時間的にも空間的にも非常にコストが高い．ここでは特に，文書データが逐次的に増加していく場合においての問題を考える．逐次的に増加していく文書データから独立な話題を得るためには，文書データが増加する度に全てのデータを用いて独立話題分析を適用する必要がある．そこで本報告では，データが増加する度に全ての文書データを用いるのではなく，増加したデータだけを用いる，データ追加に基づく独立話題分析を提案する．

以下，2 章で独立話題分析について簡単に紹介し，3 章ではデータ追加に基づく独立話題分析について提案する．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展望について述べる．

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 99, 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．

独立話題分析では，主に信号処理の分野で近年注目されている独立成分分析 (Independent Component Analysis; ICA)[Hyvärinen 00, 村田 05] を用いて話題を抽出する．独立成分分析とは，入力信号の統計的な性質を利用して異なる特性を持つ信号を分離・抽出する信号処理あるいは多変量解析の問題として定式化されている．

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し，次のように定義する $P(w) \equiv \sum_{d}^{n} a_{d,w} / \sum_{d,w}^{n,m} a_{d,w}$．ここで $a_{d,w}$ は行列 $\mathbf{A}$ の d 行 w 列の要素である．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．この時，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[Hyvärinen 99] 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．

例えば，Los Angeles Times (LA Times) の論文データ [Karypis 02, Zhong 03] 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 つを示している．

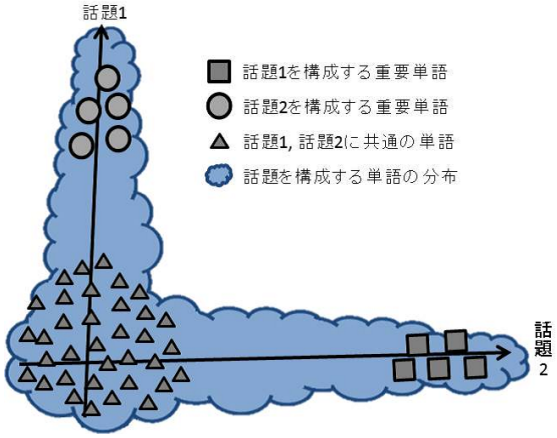

図 1: 独立話題分析のイメージ

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

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から，独立性の高い話題を得る方法である．そのため逐次的に増加していく文書データから独立な話題を得るためには，文書データが増加する度に全てのデータを用いて独立話題分析を適用する必要がある．しかし，それは時間的および空間的に非常にコストが高い．そこで本報告では，データが増加する度に全ての文書データを用いるのではなく，増加したデータだけを用いる，データ追加に基づく独立話題分析を提案する．

そこで，データが増加する度に全ての文書データを用いるのではなく，増加したデータだけを用いる，データ追加に基づく独立話題分析を提案する．

3. データ追加に基づく独立話題分析

本章では，データが追加される度に全てのデータを用いるのではなく，直前までに求めた独立な話題と，追加されたデータを用いて新たな独立な話題を求めることを考える．ここで，独立話題分析によって得られた話題の独立性の指標は，2. 章でも述べたように尖度を用いている．これは多くのデータの中から正規分布していないデータを見つけ出し，そのデータに基づいて独立性の高い話題を求めている．つまり求めた独立性の高い話題は，正規分布している多くのデータから，離れたデータであると言える．そのため独立な話題だけでは，直前までのデータを表すことは難しいと考えられる．そこで，データが追加される前に求めた独立な話題と複数個の正規分布しているデータの両方を増加してくるデータに追加して，独立性の高い

話題を求める方法を提案する．提案する方法のアルゴリズムを以下に示す．

I. 前章の独立話題分析を実行し，任意の数 k の独立性の高い話題を得る．

II. 独立な話題と正規分布している文書データを選択する．

(a) ステップ I. で得られた各話題で，文書の話題での重要度 $\mathbf{U}$ の絶対値が小さい文書データ複数個 ℓ（合計 $\ell \times k$ 個）を抜き出す．文書データに重複がある場合，重複を削除する．この $\mathbf{U}$ の絶対値が小さいデータが，多くの文書データの中の正規分布している文書データである．

(b) ステップ I. で得られた各話題の回転行列 $\mathbf{R}$ を $\mathbf{R}^{(old)}$ とする $\mathbf{R} = \mathbf{R}^{(old)}$．

(c) ステップ IIa. およびステップ IIb. 以外の文書データは削除する．

III. 追加する新しい文書データを準備する．

(a) ステップ I. と同数の数 n の追加する新しい文書データを用意する．

(b) ステップ IIIa. で用意した新しい文書データに，ステップ IIa. で抜き出した文書データ $\ell \times k$ を追加する．

IV. ステップ III. で準備した文書データを用いて，独立話題分析を実行する．

(a) 3. 章のステップ 1. とステップ 2.，およびステップ 3. とステップ 4. を順番に行う．

(b) 各話題の独立性最大化ステップ 5. を行うが，ステップ 5b で，回転行列 $\mathbf{R}$ の t 列目 $\mathbf{r}_t$ に代入する値を，ステップ IIb. で得られた $\mathbf{r}_t^{(old)}$ とする $\mathbf{r}_t = \mathbf{r}_t^{(old)}$．ここで $\mathbf{r}_t^{(old)}$ は $\mathbf{R}^{(old)}$ の t 列目のベクトルである．

(c) 得られた回転行列 $\mathbf{R}$ を用いて，${}^*\mathbf{V}$ および ${}^*\mathbf{U}$ を得る．

V. データが追加された後での独立性の高い話題が得られる．

上記のステップ II. からステップ V. を，データが追加される度に繰り返すことにより，新しい独立性の高い話題を得ることができる．

4. 実験および考察

前章で提案したデータ追加に基づく独立話題分析を評価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指標

実験では，Los Angeles Times（以下，LA Times）の新聞データ [Karypis 02, Zhong 03] を用いる．これは，文書数 6279，単語数 31472 と比較的大きな文書データである．このデータをランダムに 13 分割（1 回の文書データ数は 483）して，1 回ずつデータを追加する度に提案手法を用いて独立な話題を求めていく．また，話題の数は 6 個である．

評価は，提案した手法によって得られた話題が全てのデータを用いて独立話題分析を適用して得られた独立性の高い話題に，どの程度近づけることができたのかで行う，その評価指標には，コサイン相関の絶対値を用いる．例えば，話題 i ($i \in \{1, \ldots, k\}$) の $\mathbf{v}_i = (v_{1,i}, \ldots, v_{m,i})^{\mathrm{T}}$ と話題 j ($j \in \{1, \ldots, k\}$) の $\mathbf{v}_j = (v_{1,j}, \ldots, v_{m,j})^{\mathrm{T}}$ とのコサイン相関は次のように定義される．

$$CosDist_{ij} = \frac{\mathbf{v}_i^{\mathrm{T}} \cdot \mathbf{v}_j}{\sqrt{(\mathbf{v}_i^{\mathrm{T}}\mathbf{v}_i)(\mathbf{v}_j^{\mathrm{T}}\mathbf{v}_j)}}$$

この値の絶対値が大きければ，話題 i と話題 j は近い話題である．実験では，提案手法によって得られた各話題と全てのデータを用いて独立話題分析を適用して得られた各話題とのコサイン相関の絶対値の総和で評価する．この値が大きければ，2 つの方法によって得られた話題間は近いことを示している．また，今回の実験では話題の数は 6 であるため，この値の最大値は 6 となる．したがって，この値が 6 の時，提案手法によった得られた各話題と全てのデータを用いて独立話題分析を適用して得られた各話題は，全く同じ話題であるということを示している．

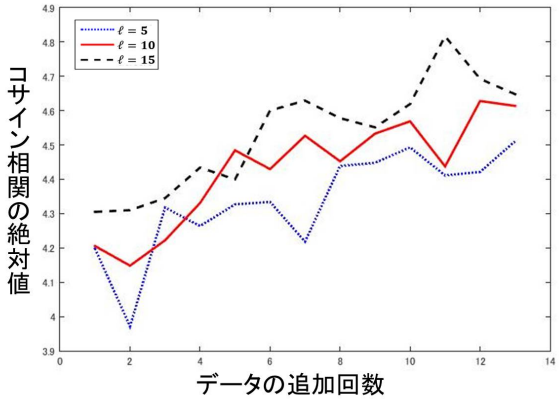

図 2: コサイン相関の絶対値とデータの追加回数

4.2 結果と考察

実験では LA Times のデータをランダムに 13 分割（1 回の文書データ数は 483）し，1 回ずつデータを追加して行った．ランダムにデータを分割して追加するため，以下では 10 回実験を行いその平均値を示す．データを追加する度に提案手法により得られた独立な話題と，全てのデータを用いて得られる独立な話題との，コサイン相関の絶対値の総和の変化を図 2 に示す．図 2 は，横軸に提案手法にデータを追加した回数を，縦軸に全てのデータを用いて得られた独立話題とのコサイン相関の絶対値をプロットしたものである．また，3. 章のステップ IIa. での，各話題での $\mathbf{U}$ の絶対値が小さいデータの個数 ℓ について，$\ell = 5$ 個の時，$\ell = 10$ 個の時，$\ell = 15$ 個の時で比較を行った．青の点線が $\ell = 5$，赤の直線が $\ell = 10$，黒の破線が $\ell = 15$ の時の結果を示している．図 2 を見ると，ℓ がいずれの場合も，回数を重ねる度に徐々に大きな値となっていくことが分かる．このことから，提案する手法はデータ追加の回数を重ねる毎に，全てのデータを用いて得られた独立話題に近づいていることが示せた．データの追加回数によって値が上下しているのは，求めた独立な話題が，直前に追加したデータの影響を強く受けてしまっているためだと考えられる．今後は，追加される前の正規分布している文書データに重みをかけるなどして，追加するデータの影響を小さくする方法を考える必要がある．また ℓ の値が大きいほど良い結果となっているが，

表 2: 提案手法 ($\ell = 10$) を LA Times に適用して得られた 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	league
2	israel	plo	palestinian	israe	arab
3	aleen	macmin	art	entertain	report
4	tower	bush	senat	committe	budget
5	compani	million	billion	earn	sale
6	polic	counti	car	arrest	kill

計算時間も増加することが分かっている．ℓ の値と計算時間とのトレードオフについて今後考える必要がある．

次に提案手法で得られた独立な各話題の重要単語を見る．表 2 に，提案手法で $\ell = 10$ の時，13 回データを追加して得られた 6 個の話題を構成する重要単語を示す．この時のコサイン相関の絶対値の総和は 4.919 であった．表 2 と表 1 を比較すると，話題 3 の単語は全て同じで，話題 1 や話題 6 は複数個の同じ単語で構成されていることが分かる．一方で表 2 の話題 2 の単語は，表 1 の話題 2 が示す「Foreign」と同じ話題を示していると考えられるが，話題を構成する重要単語は大きく異なっている．話題 4 や話題 5 も同様であることが分かる．構成する重要単語が異なる理由も，コサイン相関の絶対値が上下する理由と同じで，求めた独立な話題に対して，最後に追加した文書データの影響が大きすぎるためだと考えられる．今後の課題として，最後に追加する文書データの影響を小さくするために，データを追加する前に求めた独立な話題と正規分布している文書データに重みなどをつける，アルゴリズムの改善を考えている．

5. おわりに

本論文では，データが増加する度に全ての文書データを用いるのではなく，増加したデータだけを用いて話題を求める，データ追加に基づく独立話題分析を提案した．また，提案した方法をベンチマークデータに適用し，提案手法によって得られる話題は，データの追加を繰り返すことにより，全てのデータを用いて得られた話題に近づけることができることが示せた．

今後の課題として，提案手法を他のベンチマークデータにも適用し，その有効性を示したいと考えている．また，提案手法によって得られる話題が，全てのデータを用いて得られる独立により近い話題となるように，提案手法のアルゴリズムの改善を考えている．

参考文献

[Blei 03] Blei, David M. and Ng, Andrew Y. and Jordan, Michael I.: Latent Dirichlet Allocation, The Journal of Machine Learning Research, Vol.3, pp.993–1022 (2003).

[Deerwester 90] Scott Deerwester and Susan T Dumais and George W Furnas and Thomas K landauer and Richard Harshman: Indexing by latent semantic analysis, Journal of the American Society of Information Science, Vol.41, No.6, pp.391–407 (1990).

[Hofmann 99] Hofmann, Thomas: Probabilistic Latent Semantic Analysis, Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99), pp. 289–296 (1999).

[Hyvärinen 99] Aapo Hyvärinen: Fast and Robust Fixed-Point Algorithms for Independent Component Analysis, IEEE Trans. on Neural Networks, Vol.10, No.3 (1999).

[Hyvärinen 00] Aapo Hyvärinen and Erkki Oja: Independent component analysis: algorithms and applications, Neural Networks, Vol.13, pp.411-430 (2000).

[Hyvärinen 01] Aapo Hyvärinen and Juha Karhunen and Erkki Oja: Independent Component Analysis, John Wiley & Sons (2001).

[Karypis 02] George Karypis: CLUTO - A Clustering Toolkit, http://glaros.dtc.umn.edu/gkhome/views/cluto/ , Department of Computer Science and Engineering, University of Minnesota (2002).

[Zhong 03] 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 99] 篠原 靖志: 独立話題分析－独立性最大化による特徴的話題の抽出, 信学技法, OFS99-14 (1999).

[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).

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