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

ユーザ制約付き独立話題分析 -Separate Link 制約における制約の簡略化-

User constrained independent topic analysis
-Simplification of constraint on separation link constraint-

山本 健太 *1　西垣 貴央 *1　小野田 崇 *1
Kenta Yamamoto　Takahiro Nishigaki　Takashi Onoda

*1青山学院大学理工学部経営システム工学科
Department of Industrial and Systems Engineering，School of Science and Engineering，Aoyama Gakuin University

In this paper, we deals with topic extraction of text mining . Independent topic analysis is a method for extracting mutually independent topics from the text data by using the independent component analysis . User-constrained independent topic analysis is a method of extracting topics satisfying the user's request from the topic of independent topic analysis , and Merge Link constraint and Separate Link constraintare are proposed . We focus on Separate Link constraint that separates one of the topics into two topics . In Separate Link constraint, there is a problem that the burden on the user of selection of the word to be separated is large. In this paper, we propose method that can automate separation word selection and separate topics with high independence . We conducted an evaluation experiment with three benchmark data . We evaluated the independence of topics obtained by the proposed method and obtained the results showing its effectiveness .

1. はじめに

情報技術の発展により，現代は大量の文書データを容易に蓄積，配信できるようになった．また電子ニュースなど継続的に文書データが生成するサービスも盛んになり，日々新たな文書データが生成され続けている．このような利用可能な文書データが増え続けることに対し，人間の情報処理能力は変わらない．検索エンジンの発展により，わずかな入力で大量の文書データを得ることができるようになったが，全ての文書データを確認することは困難である．そのため，蓄積された文書データから有益な情報を抽出するテキストマイニングの研究が盛んに行われている．本論文ではテキストマイニングの課題の 1 つである話題抽出に着目する．ここでの話題とは，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である．現在では，たくさんの話題抽出方法の研究が行われており，その中でもユーザの求める話題を得る方法が必要になってきている．本論文では話題間の関係に着目した話題抽出方法の 1 つで，独立性が高い話題抽出が可能な独立話題分析 [篠原 00] に着目した．

2. 制約付き独立話題分析

独立話題分析は，主に信号処理の分野で注目されている独立成分分析 (Independent Component Analysis; ICA) [Aapo05] 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多く定式化されている．独立話題分析では 3 つの共通変数がある．話題インデックス $t \in (1,\ldots,k)$，文書インデックス $d \in (1,\ldots,n)$，単語インデックス $w \in (1,\ldots,m)$ がある．次に独立話題分析の諸概念を記述する．$\mathbf{V}$ 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．また $\mathbf{v}_t$ は，行列 $\mathbf{V}$ の t 列目のベクトル $\mathbf{v}_t = (v_{1,t},\ldots,v_{m,t})^{\mathrm{T}}$ を表し，$\mathbf{v}_w^{\mathrm{T}}$ は，行列 $\mathbf{V}$ の w 行目のベクトル $\mathbf{v}_w = (v_{w,1},\ldots,v_{w,k})$ の転置を表す．$\mathbf{U}$ 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．話題間の独立性を評価する指標には高次統計量の尖度を使用する．尖度を使用した話題の単語集中度の定義は以下のようになる．

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

$v_{w,t}$ は行列 $\mathbf{V}$ の w 行 t 列の要素である．$P(w)$ は単語 w の全文書中での出現確率を示し，定義は以下のようになる．

$$P(w) \equiv \frac{\sum_{d}^{n} a_{d,w}}{\sum_{d,w}^{n,m} a_{d,w}}$$

$a_{d,w}$ は行列 $\mathbf{A}$ の d 行 w 列の要素である．話題の単語集中度を用いることで，単語や文書の重要度を 0 に近づけることができる．少数の重要度の大きい単語や文書で話題を表現することができる．独立話題分析は，文書データから話題の単語集中度が最大となり，各話題の独立性が最大となる ${}^*\mathbf{V}$ を回転行列 $\mathbf{R}$ を用いて求める．また独立話題分析で求めたい話題数 k は人が与える．

独立話題分析による話題抽出は，話題間の独立性にのみ着目している．したがって，独立話題分析により得られた話題がユーザの求める話題と異なる場合がある．そのため，独立話題分析によって得られた話題に対し制約を加え，ユーザの要求を満たした話題を生成するユーザ制約付き独立話題分析 [西垣 16] の研究が行われている．ユーザ制約付き独立話題分析は，2 つの制約がある．1 つ目は 2 個の話題を統合する Merge Link 制約，2 つ目は 1 個の話題を 2 個に分離する Separate Link 制約がある．本論文では，1 個の話題を 2 個に分離する Separate Link 制約に着目する．

2.1 Separate Link 制約

Separate Link 制約を満たしてかつ独立性の高い話題を求める Separate Link 制約付き独立話題分析について考える． 各話題は単語の重要度を示す行列 ${}^*\mathbf{V}$ によって表現されている．ユーザはある 1 個の話題から ${}^*\mathbf{V}$ の各要素の絶対値を取った時に値が大きいものの中から，分離したい 2 個の単語（単語 p と単語 q）を選択する．分離したい単語を分離単語とよぶ．そして，選択された分離単語が重要単語となる独立性の高い話題を 2 個（話題 x と話題 y）生成する．具体的には，${}^*\mathbf{V}$ のベク

連絡先: 山本健太，青山学院大学理工学部
経営システム工学科，a5714104@aoyama.jp

トルである $\mathbf{v}_p^{\mathrm{T}}$ と $\mathbf{v}_q^{\mathrm{T}}$ で，要素の絶対値を考えた時，それぞれ話題 x と話題 y で最大値をとるように話題を生成する．制約とは関係のない話題は，はじめに得られた話題から可能な限り変化してほしくないため，制約を満たす話題 x を求めてから更新し，最後に話題 y を求める．Separate Link 制約付き独立話題分析はユーザが分離したい話題と 2 個の分離単語を選択し，その分離単語が重要単語となる独立性の高い話題を抽出する方法である．

2.2 研究課題と目的

ユーザ制約付き独立話題分析の Serprate Link 制約付き独立話題分析では，制約実行時に分離したい話題の選択と 2 個の分離単語の選択が必要がある．特に 2 個の分離単語の選択はユーザの確認量がとても多くなり，ユーザの負担が大きいことが課題である．以上の課題点から本論文では Separate Link 制約実行時のユーザの負担の少ない話題抽出方法を考える．提案手法としてユーザが分離単語を選択せずに，より独立性の高い話題が得られる方法が必要である．

3. 提案手法

Separate Link 制約付き独立話題分析の課題点を解決する話題抽出方法を考える．既存手法の Separate Link 制約付き独立話題析分析は分離したい話題の中でユーザが分離単語を選択し，話題を分離する方法である．提案する話題抽出方法は Separate Link 制約実行時の分離単語の選択によるユーザの負担が少ない方法を考える．また，制約に関係のない話題は制約を与える前の話題から大きく変化しないものと仮定する．本論文の提案手法はユーザの負担を減らすため分離単語選択の自動化を行う．分離単語選択の自動化は，制約実行後の話題間の独立性が高くなるように自動で単語を選択する．Separate Link 制約を満たし，ユーザが分離したい話題を選択するだけで独立性高く話題を分離可能な話題抽出方法を提案する．

3.1 アルゴリズム

提案手法を用いた Separate Link 制約付き独立話題分析のアルゴリズムを以下に述べる．アルゴリズムは 12 個のステップで構成されている．

1. 独立話題分析で任意の数 k の独立な話題を得る．ただし，独立話題分析において，$k+1$ 個の成分を抜き出し，$(k+1)\times m$ の行列 $\mathbf{X}$ を得る．また，回転行列 $\mathbf{R}$ の初期値を $(k+1)\times(k+1)$ の零行列として k 個の話題を求めた．なお，$k+1$ 番目の列は全て零となっているが，これは回転行列を正方行列にするための処置であり，独立話題分析の特性上問題ない．
2. Separate Link 制約を与える話題 z ($z \in \{1,\ldots,k\}$) をユーザが選択する．
3. 文書を話題ごとに分類する．各文書がどの話題に分類されるかは，話題に対する文書の重要度 $\mathbf{U}$ を用いる．各文書の重要度の最も大きい話題に分類する．

   $$C_{\mathrm{t}} = \{a_d | \arg\max_{1\le t\le k} u_{d,t}\},\ C_{\mathrm{t}} \in \{1,\ldots,k\}$$

   Separate Link 制約を与える話題 z に分類された以下の式を満たした文書データを $\mathbf{F}$ とする．また文書データを $\mathbf{F}$ の文書数を h とする．

   $$C_{\mathrm{z}} = \{a_d\}$$

4. Separate Link 制約を与える話題 z に分類された文書データ $\mathbf{F}$ に再び独立話題分析を行い特徴抽出を行う．独立話題数は 2 とし，2 つの独立な話題を求める．ここで求めた話題における単語の重要度を $\mathbf{V}^{\dagger}$ とする．
5. 求めた各話題における単語の重要度 $\mathbf{V}^{\dagger}$ の最も大きい単語を制約での分離単語 p, q とする．

   $$\text{分離単語 } p = \arg\max_{1\le w\le m} v^{\dagger}_{w,1}\ (p \in \{1,\ldots,m\})$$

   $$\text{分離単語 } q = \arg\max_{1\le w\le m} v^{\dagger}_{w,2}\ (q \in \{1,\ldots,m\})$$

   ここで ${\mathbf{v}^{\dagger}}_t$ は，行列 $\mathbf{V}^{\dagger}$ の t 列目のベクトル ${\mathbf{v}^{\dagger}}_t = (v^{\dagger}_{1,t},\ldots,v^{\dagger}_{m,t})^{\mathrm{T}}$ を表す．
6. 単語 p が重要単語となる話題 x を生成する回転ベクトルを次の式で設定する．

   $$\mathbf{c} = (\mathbf{u}^{\mathrm{T}}_{*\alpha}).^2, \quad \mathbf{c} = \mathbf{c}/\|\mathbf{c}\|$$

   $$\alpha = \arg\max_{1\le d\le h} \mathbf{f}_{d,p}, \quad \alpha \in \{1,\ldots,h\}$$

   ここで $\mathbf{u}^{\mathrm{T}}_{\alpha}$ は，行列 $\mathbf{U}$ の α 行目のベクトルの転置を表す．また $\alpha = \arg\max_{1\le d\le n} \mathbf{f}_{d,p}$ は，行列 $\mathbf{F}$ での単語 p を示す列ベクトル $\mathbf{f}_p$ の各要素の値が最も大きい文書番号（行番号）を α としている．$^*\alpha$ は全文書データ内での α の文書番号（行番号）を表す．
7. Separate Link 制約を満たして独立な話題の回転行列を $\ddot{\mathbf{R}}$ として，その最初の列にステップ 6 で求めた $\mathbf{c}$ を代入する．
8. ステップ 1 で得た回転行列 $\mathbf{R}$ の z 列目を抜き出した $(k-1)\times k$ の行列を独立話題分析での $\mathbf{e}_t$ の代わりとして，独立話題分析と同様に $2 \le t \le k$ までの新しい回転行列 $\ddot{\mathbf{R}}$ を求める．
9. 最後にステップ 6 と同様に，単語 q が重要単語となる話題 y を生成する回転ベクトルを次のように表す．

   $$\mathbf{g} = (\mathbf{u}^{\mathrm{T}}_{*\beta}).^2, \quad \mathbf{g} = \mathbf{g}/\|\mathbf{g}\|$$

   $$\beta = \arg\max_{1\le d\le h} \mathbf{f}_{d,q}, \quad \beta \in \{1,\ldots,h\}$$

10. ステップ 9 で得た $\mathbf{g}$ を次の回転行列化を行う．

    $$\mathbf{g} = \mathbf{g} - \ddot{\mathbf{R}}\ddot{\mathbf{R}}^{\mathrm{T}}\mathbf{g}, \quad \mathbf{g} = \mathbf{g}/\|\mathbf{g}\|$$

11. $\mathbf{g}$ を $\ddot{\mathbf{R}}$ の最後の列に代入し，その時の $\ddot{\mathbf{R}}$ が Separate Link 制約を満たす独立な話題の回転行列である．
12. 新たな $k+1$ 個の独立な話題を得る．

これらのステップ 2 からステップ 12 を複数回繰り返すことで，話題を 1 個ずつ増やしていくことが可能である．この方法で Separate Link 制約を満たし，独立性の高い話題を求める．

3.2 既存手法との比較

既存手法と提案手法の特徴の比較を記述する．既存手法はまず独立話題分析の結果からユーザが分離したい話題を選択する．次に分離したい話題の重要単語の中からユーザが 2 個の分離単語（単語 p と単語 q）を選択する．$^*\mathbf{V}$ のベクトルである $\mathbf{v}_p^{\mathrm{T}}$ と $\mathbf{v}_q^{\mathrm{T}}$ で，要素の絶対値を考えた時，それぞれの分離話題で最大値をとるような話題を抽出する話題抽出方法である．既存手法はユーザの分離したい単語を 2 個の話題に分離可能な話題抽出方法である．

対して提案手法は，まず独立話題分析の結果からユーザが分離したい話題を選択する．次に分離したい話題の重要単語の

中から 2 個の分離単語（単語 p と単語 q）を制約実行後の話題間の独立性が高くなるように分離単語を自動で抽出する．既存手法同様に ${}^*\mathbf{V}$ のベクトルである $\mathbf{v}_p^{\mathrm{T}}$ と $\mathbf{v}_q^{\mathrm{T}}$ で，要素の絶対値を考えた時，それぞれの分離話題で最大値をとるような話題を抽出する話題抽出方法である．提案手法は，分離したい話題を独立性高く 2 個の話題に分離可能な話題抽出方法である．

図 1 は既存手法の分離単語選択の一例を表している．図 2 は提案手法の分離単語選択を表している．各図では話題 1 に制約を与え，話題の分離を行うと仮定する．話題 1 を構成する単語を赤円，話題 2 を構成する単語を青円，多くの話題に共通する単語を四角で表現している．Separate Link 制約では分離単語をユーザが重要単語の中から自由に選択することができる．自由に分離単語を選択できるため，得られる話題は一意には定まらない．また図 1 のように単語間の関係性が高い単語も分離単語に選択される場合もあり，話題間の独立性も分離単語の選択で大きく変化する．既存手法はユーザが分離したい単語を選択し分離することが可能な制約である．しかしユーザは分離単語選択の負担が大きく，その分離単語選択で抽出される話題や話題間の独立性が変化する．対して，提案手法ではユーザが分離単語を選択する必要はなく，図 2 のように制約実行後の話題間の独立性が高くなるよう分離単語を自動で決定する．また提案手法では分離単語と得られる話題は一意に定まる．そして得られる話題が定まるため，話題間の独立性も一意に定まる．以上の比較を表 1 にまとめる．

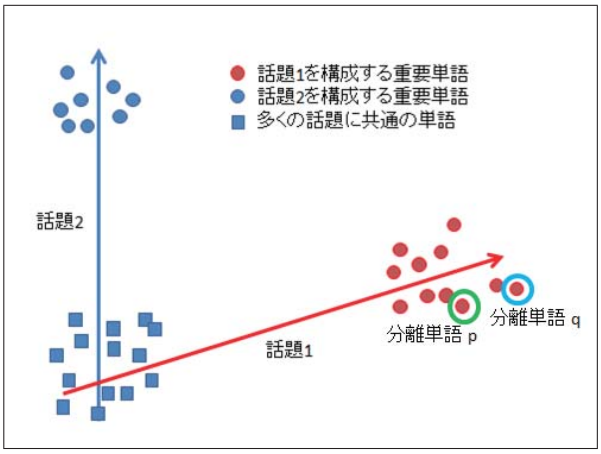

図 1: 既存手法の分離単語例

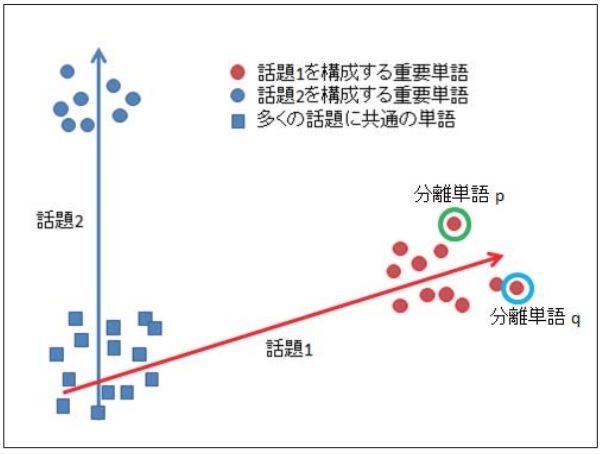

図 2: 提案手法の分離単語例

表 1: 制約の比較

手法	既存手法	提案手法
話題選択	行う	行う
単語選択	行う	行わない
話題	分離単語選択で変化	一意に定まる
独立性	分離単語選択で大きく変化	独立性高く定まる

4. 評価実験

評価実験では，Separate Link 制約を満たしているかの充足性および，提案手法によって得られた話題間の独立性の評価する．

4.1 実験条件

実験データは 3 つのベンチマークデータを用いた．

- $LosAngelsTimes(LATimes)$ の新聞データで文書数は 6279，単語数は 31472 の文書データ [la12][CLUTO]
- $DAILY\,KOSblog(KOS)$ のブログデータで文書数は 3430，単語数は 6906 の文書データ [UCIrep]
- $NeuralInformationProcessingSystems(NIPS)$ の論文データで文書数は 1500，単語数は 12419 の文書データ [UCIrep]

実験の評価は Separate Link 制約を満たしているかの充足性および，提案手法によって得られた話題間の独立性の評価する．まず Separate Link 制約の充足性の評価では，各話題における単語の重要度を示す行列 $\mathbf{V}$ を用いる．2 つの分離単語の各話題における重要度の値から，それぞれの分離話題の重要度が最も大きいことを確認する．

提案手法によって得られた話題間の独立性の評価は，制約により抽出された 2 個の分離話題間の独立性と制約に関係ない話題を含めた全話題間の独立性を評価する．独立性の評価は話題間の相互情報量を用いる．話題間の相互情報量が 0 の場合，2 つの話題は完全に独立していることを意味し，相互情報量が小さいほど話題間の独立性は高くなる．相互情報量は独立話題分析によって得られた $\mathbf{V}$ の値を用いて求める．提案手法は分離単語と得られる話題は一意に定まるが，既存手法はユーザが分離単語を選択するため，得られる話題は一意には定まらない．そこで既存手法との比較では，分離単語を重要単語の中からランダムで選択した 20 回の平均を用いて行う．

4.2 実験結果，考察

実験結果の一例として $LATimes$ に話題数 7 で独立話題分析を行った結果の話題 6 に評価実験を行った結果を示す．$LATimes$ に話題数 7 で独立話題分析を行った結果を表 2 に示す．まず，提案手法は分離したい話題の文書を抽出し，再度その文書に独立話題分析を行う．分離したい話題 6 の文書に対し，独立話題分析を行なった結果を表 3 に示す．分離したい話題 6 を構成する 2 つの話題が抽出された．提案手法での分離単語は抽出された表 3 の各話題で最も重要度の大きい bank と price に決定された．次に bank と price を分離単語として話題 6 を分離した結果を表 4 に示す．

次に提案手法により得られた話題の評価を行う．まず，Separate Link 制約の充足性の評価として各話題における分離単語 bank と price の重要度を表 5 に示す．表 5 よりそれぞれの分離話題での重要度が最も大きいことから Separate Link 制約の充足性が確認できる．次に提案手法によって得られた話題間の独立性の評価を表 6 に示す．分離話題間の独立性，全話題

間の独立性ともに提案手法により得られた話題間の独立性が高く，提案手法の有効性が確認できる．*LATimes* に異なる話題数，話題の評価実験の結果，*Nips*, *Kos* への適応結果も同様に提案手法によって得られた話題間の独立性が高く，提案手法の有効性が確認できた．

比較手法の独立性が低くなった理由としては，重要単語の中からランダムに分離単語を選択することで分離単語間の関係性が高い単語のペアが選択される場合があることが考えられる．また制約に関係のない話題の重要単語と関係性が高い単語が分離単語として選択される場合があることも理由の 1 つであると考えらえる． 単語間の関係性が高い分離単語を用いて話題を分離することで，分離話題間の相互情報量や分離話題と制約に関係のない話題間の相互情報量が高くなり独立性が低くなる．対して提案手法は分離したい話題がどのような話題で構成されているかを分析し，2 個の分離単語を決定することで独立性が高くなるように単語を選択することができる．その結果提案手法は比較手法と比べ独立性が高い話題抽出ができたと考えられる．

表 2: *LATimes* の新聞データに話題数 7 で独立話題分析による話題抽出結果

話題	重要度が高い単語				
	$w=1$	$w=2$	$w=3$	$w=4$	$w=5$
1	million	earn	quarter	revenu	net
2	scor	game	lead	rebound	league
3	soviet	afghanistan	israel	foreign	militari
4	macmin	aleen	art	entertain	report
5	polic	bush	counti	car	arrest
6	stock	bank	price	market	rate
7	game	team	player	coach	bowl

表 3: 独立話題分析結果表 2 の分離したい話題 6 の文書に独立話題分析実行結果

話題	重要度が高い単語				
	$w=1$	$w=2$	$w=3$	$w=4$	$w=5$
1	bank	insur	loan	sav	deposit
2	price	bond	index	stock	rate

表 4: 独立話題分析結果表 2 の話題 6 に bank と price を分離単語とし，提案手法による話題抽出結果

話題	重要度が高い単語				
	$w=1$	$w=2$	$w=3$	$w=4$	$w=5$
1	stock	polic	bank	price	market
2	million	earn	quarter	revenu	net
3	scor	game	lead	rebound	league
4	soviet	afghanistan	israel	foreign	militari
5	macmin	aleen	art	entertain	report
6	polic	bush	counti	car	arrest
7	game	team	player	coach	bowl
8	earn	quarter	stock	price	rate

5. まとめ

本論文では，独立性の高い話題抽出方法である独立話題分析にユーザ制約を加えたユーザ制約付き独立話題分析を扱った．そして，ユーザ制約付き独立話題分析の Separate Link 制約付き独立話題分析に着目した．Separate Link 制約は 1 個の話題を 2 個の話題に分離する制約である．Separate Link 制約付き独立話題分析にはユーザの分離単語選択の負担が大きい課題がある．提案手法は分離単語選択の自動化により Separate Link 制約実行時のユーザの負担を軽減し，独立性の高い話題を抽出する方法を考案した．分離単語の自動化ではまず，分離したい話題にどのような話題が含まれているかを独立話題分析により再度抽出した．そして再度独立話題分析した結果から自動で分離単語を選択する方法を用いた．実験では提案手法を 3 つのベンチマークデータに適応させた．実験結果から提案手法は，ユーザの負担を軽減した上で独立性の高く話題を分離可能な手法であると示された．今後は提案手法で得られる話題の独立性の高さの理論的証明が必要であると考えられる．また分離できる話題数を 3 個，4 個と増やすことで汎用性が高まると考えている．

表 5: 提案手法による話題抽出結果表 4 での分離単語 bank と price の各話題の重要度 $|v_{w,t}|$

	$w=$ bank	$w=$ price
$v_{w,\text{ 話題 }1}$	**34.243**	22.963
$v_{w,\text{ 話題 }2}$	1.028	12.963
$v_{w,\text{ 話題 }3}$	8.056	2.709
$v_{w,\text{ 話題 }4}$	8.208	0.973
$v_{w,\text{ 話題 }5}$	7.197	3.609
$v_{w,\text{ 話題 }6}$	3.790	10.826
$v_{w,\text{ 話題 }7}$	17.410	1.151
$v_{w,\text{ 話題 }8}$	25.03	**36.302**

表 6: 提案手法により得られた話題間の独立性の評価

	提案手法	比較手法 (ランダム)
分離話題間	**0.0383**	0.0680
全話題間	**2.1034**	2.4624

参考文献

[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，電力中央研究所報告，(2000)

[Aapo05] Aapo Hyvarinen， Juha Karhunen， Erkki Oja: 詳解　独立成分分析　信号解析の新しい世界，東京電機大学出版局，(2005)

[西垣 16] 西垣 貴央， 新田 克己， 小野田 崇:制約付き独立話題分析，人工知能学会論文誌，(2016)

[la12] Shi Zhong， Joydeep Ghosh:A comparative study of generative models for document clustering，Data Mining Workshop on Clustering High Dimensional Data and Its Applications，(2003)

[CLUTO] Ying Zhao， George Karypis:Evaluation of hierarchical clustering algorithms for document datasets，Conference of Information and Knowledge Management (CIKM)，(2002)

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