論文 / 著書情報
Article / Book Information

題目(和文)	制約付き独立話題分析に関する研究
Title(English)	Constrained Independent Topic Analysis
著者(和文)	西垣貴央
Author(English)	Takahiro Nishigaki
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第10541号, 授与年月日:2017年3月26日, 学位の種別:課程博士, 審査員:新田 克己,寺野 隆雄,渡邊 澄夫,小野 功,石井 秀明,小野田 崇
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第10541号, Conferred date:2017/3/26, Degree Type:Course doctor, Examiner:,,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

博 士 論 文

制約付き独立話題分析に関する研究

Constrained Independent Topic Analysis

日付：2017 年 2 月 24 日

指導教員：新田 克己 教授

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

氏名：西垣 貴央

Department of Computational Intelligence and Systems Science,
Interdisciplinary Graduate School of Science and Engineering,
Tokyo Institute of Technology
Takahiro NISHIGAKI

目次

第 1 章　序論 4

1.1　研究背景 4

1.2　研究課題および目的 8

1.3　論文構成 9

第 2 章　独立話題分析 10

2.1　概念 10

2.2　方法 12

2.3　課題 14

2.4　本章まとめ 16

第 3 章　関連研究 17

3.1　ユーザ制約付き話題抽出 17

3.2　データ追加に基づいた話題抽出 19

3.3　本章まとめ 21

第 4 章　ユーザ制約付き独立話題分析 22

4.1　新たな制約の定義 22

4.2　制約付きクラスタリングでの制約との違い 23

4.3　Merge Link 制約付き独立話題分析 23

4.3.1　目的 23

4.3.2　方法 24

4.3.3 評価方法 27
4.3.4 実験結果及び考察 29
4.4 Separate Link 制約付き独立話題分析 37
4.4.1 目的 37
4.4.2 方法 37
4.4.3 評価方法 39
4.4.4 実験結果及び考察 41
4.5 本章まとめ 49

第 5 章 データ追加に基づく独立話題分析 50
5.1 目的 50
5.2 方法 50
5.3 評価方法 53
5.4 実験結果及び考察 55
5.5 本章まとめ 71

第 6 章 結論 72

参考文献 77

第 1 章

序論

本章では, 制約付き独立話題分析に関する研究の背景, 研究課題と目的, および本論文の構成について述べる.

1.1 研究背景

近年, ノート PC やタブレットなどの計算機ハードウェアの普及により, ブログや Web ニュースなどで様々な文書データが生成されている. また, ハードディスクドライブ (HDD) などの大容量記憶媒体の性能向上や, インターネットによる情報通信技術の高度化に伴い, 膨大な量の文書データが蓄積されている. これらの膨大な量の文書データから, 有益な知識を発見・抽出するための技術であるテキストマイニングが研究されている [渡部 03]. 本論文では, テキストマイニングの課題の一つである話題抽出について取り上げる.

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のことである [佐藤 15]. この話題を抽出する方法として, Hofmann の提案した PLSA (Probabilistic Latent Semantic Analysis)[Hofmann99a] や Blei らの提案した LDA (Latent Dirichlet Allocation)[Blei03] をはじめとしたトピックモデルが研究されている. トピックモデルとは, 確率的生成モデルに着目して話題を抽出する方法である [Blei12]. トピックモデルでは話題

を潜在変数と考え，その潜在変数に依存した混合分布として単語の出現確率の分布を獲得する．またトピックモデルでは文書ごとに話題の分布を定義することにより，文書の持つ話題の偏りを自然に表現できる．多くのトピックモデルでは，文書を単語の集合 (bag-of-words) として扱う．bag-of-words とは単語の出現順序を無視したもので，多くの場合，単語の頻度や tf-idf[Salton83] の値を用いる．トピックモデルを用いることによって文書，単語と話題の関係を確率モデルで表すことが可能である．しかし，PLSA や LDA では話題間の関係，例えば相関関係や独立性には着目していない．

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，LSI (Latent Semantic Indexing)[Deerwester90] や独立話題分析 [篠原 99, 篠原 00] がある．LSI では文書データに特異値分解を適用することで，単語や文書の分散が最も大きくなるような話題を求めることができる．この方法によって，話題間に相関関係のない話題を得ることができる．相関関係のない話題とは，ある話題が増えると別のある話題も増えるといった直線的な関係がないことを示している．

相関関係のない話題を求める方法とは異なり，独立性の高い話題を求め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 [Hyv*ä*rinen01, 村田 05] を用いて話題を求めている．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．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．より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あげられる．また，独立な話題は相関関係のない話題を含むという関係が成り立っており，相関関係のない話題と独立な話題は等価ではない．ただし与えられる文書データ内の話題が正規分布している場合は，相関関係のない話題は独立な話題と同じ話題

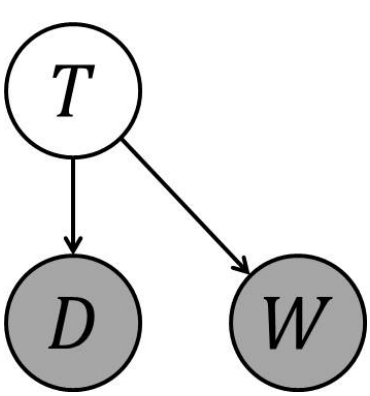
となる．しかし実際に与えられる文書データ内の話題は，必ずしも正規分布しているとは限らない．そのため話題間の相互情報量が小さい話題を得るためには，独立性の高い話題を求める必要がある．

本論文では独立性に着目して話題を抽出する方法について考える．話題の独立性を考慮した研究としてトピックモデルの一つである PLSA に独立な情報の概念を取り入れた STI-PLSA[神嶌 15] や得られる話題間の独立性が高くなるように話題を抽出する独立話題分析などが存在する．STI-PLSA とは神嶌らによって提案された，従来のトピックモデルである PLSA に独立なセンシティブ情報という概念を付与した新しいトピックモデルである．以下，本論文共通の変数として，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．また，記号の小文字はスカラー，小文字太字はベクトル，大文字太字は行列を表す．

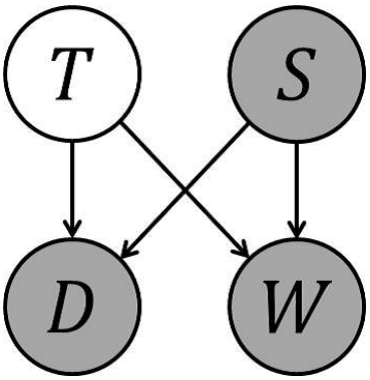
まず，PLSA[Hofmann99a] について簡単に述べる．PLSA とは確率的生成モデルに着目して文書，単語と話題の関係をモデル化したものである．数式で表すと次のようになる．

$$Pr(D, W) = \sum_{T} Pr(D|T)Pr(W|T)Pr(T)$$

ここで，D は文書を表す確率変数であり，その実現値は $d \in \{1, \ldots, n\}$ となる．W は単語を表す確率変数であり，その実現値は $w \in \{1, \ldots, m\}$ となる．また，T は話題を表す確率変数であり，実現値は $t \in \{1, \ldots, k\}$ となる．$Pr(d|t)$ と $Pr(w|t)$ はそれぞれ，話題 t に対する文書と単語の関連の強さを表すことになる．これらの値が大きいと，その話題を示す文書であったり，その話題でよく使われる単語であったりする．$Pr(T)$，$Pr(D|T)$，$Pr(W|T)$ は最尤推定により求める．この最尤推定は EM アルゴリズムを利用することで解けることが広く知られている [Hofmann99b]．また，PLSA での各変数間の依存関係をグラフィカルモデルで表すと図 1.1(a) のようになる．図 1.1 では観測変数を黒色で潜在変数を白色で表している．神嶌らはこの PLSA に話題と独立な情報を表すセンシティブ変数 S を導入した厳密ト



(a) PLSA のグラフィカルモデル



(b) STI-PLSA のグラフィカルモデル

図 1.1　PLSA と STI-PLSA のグラフィカルモデル：T は話題，D は文書，W は単語，S は独立なセンシティブ情報を示し，黒丸が観測変数，白丸が潜在変数を示す．

ピック独立 PLSA (Strictly Topic-Independent PLSA; STI-PLSA) を提案した．センシティブ変数 S は話題 T との間で条件なし独立 $S \perp\!\!\!\perp T$ を満たす．この STI-PLSA を数式で表すと次のようになる．

$$Pr(D, W, S) = Pr(S) \sum_{T} Pr(D|T, S) Pr(W|T, S) Pr(T)$$

PLSA と同様に，$Pr(T)$，$Pr(S)$，$Pr(D|T, S)$，$Pr(W|T, S)$ は最尤推定の EM アルゴリズムを利用することで解くことができる．また，STI-PLSA での各変数間の依存関係をグラフィカルモデルで表すと図 1.1(b) のようになる．この STI-PLSA では，観測変数である文書 D と単語 W，話題 T と独立なセンシティブ変数 S から，話題 T を得るものである．つまり，この方法で得られる話題間 $t \in \{1, \ldots, k\}$ の独立性は仮定のみ行っているだけで，従来のトピックモデルと同様にその仮定が成り立っているのかどうか明確ではない．

本論文では，話題の独立性でも特に話題間の独立性に着目して話題を求める方法について検討する．そこで，話題間 $t \in \{1, \ldots, k\}$ の独立性に焦点を当て，得られる話題間が独立である話題を得ることができる独立話題分析に注目する．

1.2 研究課題および目的

本論文では，独立話題分析の二つの課題について検討する．第一の課題は，独立話題分析によって得られる話題は，話題の独立性にのみ着目して求めているので，得られる話題がユーザの求める話題と異なる場合が存在するというものである．第二の課題は，独立話題分析では逐次増加するデータには適用が難しいということである．独立話題分析では独立な話題を求めるために，全てのデータを使用して話題を求めるため，新しいデータが入ってから再び独立な話題を求める場合，再び全てのデータを使用する必要がある．

これら二つの課題についての研究は，トピックモデルやクラスタリング手法においては多く行われている．しかし，話題の独立性に着目して話題を求める独立話題分析に，ユーザ制約を取り入れる方法や，増加するデータに適用できる方法についての研究は進んでいない．そこで本研究の目的は，独立話題分析において，これら二つの課題を解決する方法を提案することである．

本論文では，第一の課題に対しては，得られる話題をユーザの求める話題に近づけるため，独立話題分析にユーザ制約を取り入れることで，ユーザ制約を満たして，かつ独立性の高い話題を求める方法を提案する [西垣 16b]．第二の課題に対しては，新しいデータが入ってきても全てのデータを使用しないで独立な話題を求める事ができるように，初期データから求めた独立性の高い話題を，データが増加する度に，増加したデータの独立性に基づいて更新することで，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を提案する [西垣 16a, Nishigaki17].

ユーザ制約として与えられる情報のことを制約あるいは教師と呼ぶが，本論文では [神嶌 06] に基づいて教師ではなく制約と呼ぶ．また，使用できるデータ数に制限があると考えることで，こちらについても広義の制約であると言うことができる．

1.3　論文構成

本論文では，制約を取り入れた制約付き独立話題分析について検討を行った．本論文の構成は以下のとおりである．

2 章では，話題間が独立な話題となるように話題を抽出する手法である独立話題分析とそのアルゴリズムについて述べる．そして，独立話題分析には，得られる話題はユーザが求める話題と異なる場合が存在するという課題と，話題を抽出する際全てのデータを使用する必要があるため，逐次増加するデータに対して適用が難しいという二つの課題が存在することを説明する．

3 章関連研究では，最初にユーザ制約付き話題抽出に関する研究について紹介する．そして，本論文で提案する独立話題分析へのユーザ制約を加えることの必要性があることを示す．続いてデータが逐次増加する場合における話題の抽出に関する研究について紹介する．さらに，本論文で提案する独立話題分析を逐次増加するデータへの適用する方法の必要性を示す．

4 章では，話題間の独立性に着目して独立な話題を抽出する独立話題分析にユーザ制約を加える方法について述べる．最初に，本論文で新しく提案する二つの制約について説明する．続いて，新たに提案した制約と制約付きクラスタリングで一般的に用いられる制約との違いを述べる．さらに，それぞれの制約を満たしてかつ，独立な話題を求めるユーザ制約付き独立話題分析の方法について提案し，その提案した方法の評価実験および考察を行う．

5 章では，最初にデータ追加に基づく独立話題分析の目的を述べ，続いてその方法について述べる．さらに，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評価方法について説明し，評価実験の結果および考察を行う．

最後に，6 章で本論文をまとめる．

第 2 章

独立話題分析

本章では，話題間が独立となるように話題を抽出する方法である，篠原によって提案された独立話題分析 [篠原 00] について述べる．

2.1 概念

篠原によって提案された独立話題分析では，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)[Hyvärinen00, 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 [Hyvärinen01, 村田 05].

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．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．また $\mathbf{u}_t$ は，行列 $\mathbf{U}$ の t 列目のベクトル $\mathbf{u}_t = (u_{1,t}, \ldots, u_{n,t})^{\mathrm{T}}$ を表し，$\mathbf{u}_d^{\mathrm{T}}$ は，行列 $\mathbf{U}$ の d 行目のベクトル $\mathbf{u}_d = (u_{d,1}, \ldots, u_{d,k})$ の転置を表す．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 $\mathbf{a}_w$ は，行列 $\mathbf{A}$ の w 列目のベ

クトル $\mathbf{a}_w = (a_{1,w}, \ldots, a_{n,w})^{\mathrm{T}}$ を表し，$\mathbf{a}_d^{\mathrm{T}}$ は，行列 $\mathbf{A}$ の d 行目のベクトル $\mathbf{a}_d = (a_{d,1}, \ldots, a_{d,m})$ の転置を表す．つまり，行列 $\mathbf{V}$ と行列 $\mathbf{U}$，行列 $\mathbf{A}$ は次のように定義される．

$$\mathbf{V} = \begin{pmatrix} v_{1,1} & \cdots & v_{1,k} \\ \vdots & \ddots & \vdots \\ v_{m,1} & \cdots & v_{m,k} \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} u_{1,1} & \cdots & u_{1,k} \\ \vdots & \ddots & \vdots \\ u_{n,1} & \cdots & u_{n,k} \end{pmatrix}$$

$$\mathbf{A} = \begin{pmatrix} a_{1,1} & \cdots & a_{1,m} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,m} \end{pmatrix}$$

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．

話題の単語集中度

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

$v_{w,t}$ は行列 $\mathbf{V}$ の w 行 t 列の要素である．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．

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

$a_{d,w}$ 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．

2.2 方法

独立話題分析は，前節で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 ${}^*\mathbf{V}$ と ${}^*\mathbf{U}$ を回転行列 $\mathbf{R}$ を用いて求める．独立話題分析を最適化問題として定式化すると次のようになる．

$$\begin{aligned} &\underset{\mathbf{R}}{\text{maximize}} \quad \left| \sum_{t}^{k} \left\{ \sum_{w}^{m} \left((\mathbf{VR}).^4 P(w) \right) - 3 \left(\sum_{w}^{m} (\mathbf{VR}).^2 P(w) \right)^2 \right\} \right| \\ &\text{subject to} \quad \mathbf{R}^{\mathrm{T}}\mathbf{R} = 1, \quad \|\mathbf{R}\| = 1 \end{aligned}$$

ここで $(\mathbf{VR}).^4$ は，行列 $\mathbf{VR}$ の各要素の 4 乗を表す．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，FPICA[Hyvärinen99] と同様の方法で正規化を行い $\tilde{\mathbf{A}}$ を得る．
2. $\tilde{\mathbf{A}}$ に対して特異値分解を行い，行列 $\tilde{\mathbf{A}}$ 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 $\hat{\mathbf{U}}$，$\hat{\mathbf{S}}$，$\hat{\mathbf{V}}$ を，行列 $\hat{\mathbf{S}}$ 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 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.

$$\mathbf{r}^{(old)} = \mathbf{r}_t, \quad \mathbf{r}_t = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{r}_t).^3 - 3\mathbf{r}_t$$

$(\mathbf{X}^{\mathrm{T}}\mathbf{r}_t).^3$ 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.

$$\mathbf{r}_t = \mathbf{r}_t - \mathbf{R}\mathbf{R}^{\mathrm{T}}\mathbf{r}_t, \quad \mathbf{r}_t = \mathbf{r}_t / \|\mathbf{r}_t\|$$

（f）FPICA[Hyvärinen99] と同様の条件で収束しなければ，ステップ (5d) へ．収束すればステップ (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.

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

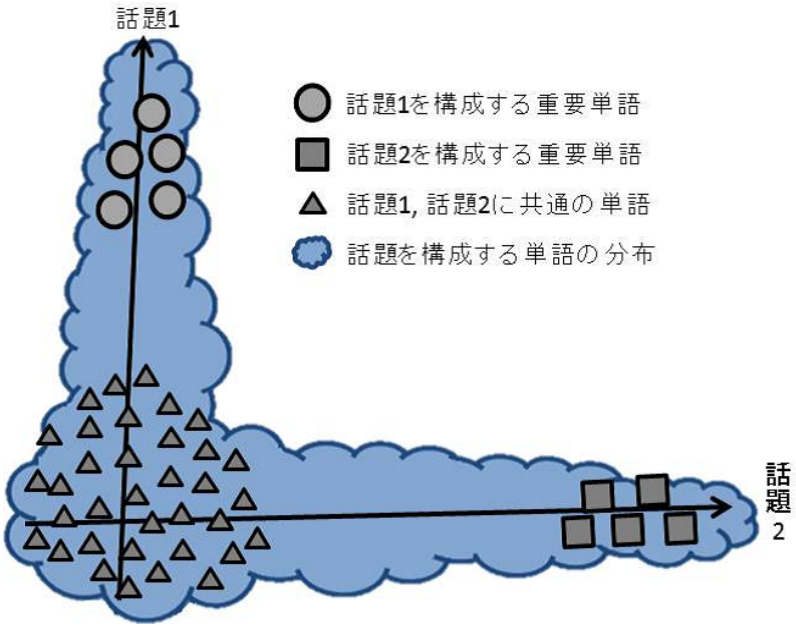


図 2.1　独立話題分析のイメージ

以上の独立話題分析によって得られる話題は，図 2.1 のようになり独立性の高い話題を得ることができる．

2.3　課題

この独立話題分析には，二つの課題が存在する．第一の課題は，独立話題分析によって得られる話題は，話題の独立性にのみ着目して求めているので，得られる話題がユーザの求める話題と異なる場合が存在するというものである．例えば，Los Angeles Times (LA Times) の新聞データ [Zhao02, Zhong03] に対して話題数 7 で独立話題分析を行うと表 2.1 のように話題が抽出できる．表 2.1 の話題を構成する重要単語を見ると，話題 1 は「Revenue」, 話題 2 は「Soccer」, 話題 3 は「Foreign」, 話題 4 は「Entertainment」, 話題 5 は「Affairs」，話題 6 は「Stock」，話題 7 は「Team, Player」の内容を表していると考えられる．表 2.1 の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 **V** の各話題（各列）において要素の絶対値が大きいもの五つを示している．

表 2.1　LA Times に独立話題分析を適用して得られた 7 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	million	earn	quarter	revenu	net
2	scor	game	lead	rebound	league
3	soviet	afghanistan	israel	foreign	militari
4	aleen	macmin	art	entertain	report
5	polic	bush	counti	car	arrest
6	stock	bank	price	market	rate
7	game	team	player	coach	bowl

この時ユーザが，話題 1 と話題 6 はどちらも金融の話題を表しているので，1 個の話題「Finance」として，合計 6 個の話題を得たいと考える場合がある．また，反対に話題 5 の「Affairs」を分離させて，「Los Angeles」の話題と「President」の話題を抽出して，合計 8 個の話題を得たいと考える場合がある．このように 2 個の話題を 1 個の話題に統合する，あるいは 1 個の話題を 2 個の話題として分離するといったユーザ制約を取り入れた新たな話題を得る方法が必要となる．しかし，このようなユーザ制約を満たす独立話題分析については提案されていない．そこで 4 章では，独立話題分析にユーザ制約を取り入れ，ユーザの制約を満たし，かつ独立性の高い話題を求める，ユーザ制約付き独立話題分析を提案する．

第二の課題は，独立話題分析では，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ということである．なぜなら，独立話題分析は現在所持している全てのデータを使用して，独立な話題を抽出する方法であるからである．図 2.2 にその様子を示す．しかし，増加するデータに独立話題分析を適用する方法については提案されていない．そこで本研究の目的は，これら二つの課題を解決する方法を提案することである．そこで 5 章では，数が増加していくデータに独立話題分析を適用できる，データ追加に基づく独立話題分析を提案する．

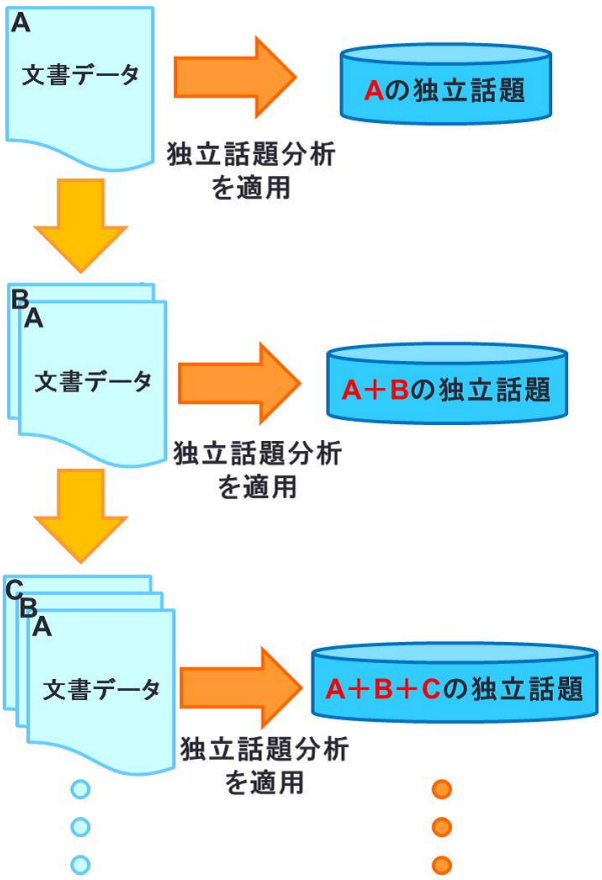


図 2.2　独立話題分析の逐次増加するデータへの適用の課題のイメージ

2.4　本章まとめ

本章では，話題間が独立となるように話題を抽出する手法である独立話題分析とそのアルゴリズム，および独立話題分析が持つ二つの課題について述べた．第一の課題は，独立話題分析によって得られる話題がユーザの求める話題と異なる場合が存在するというものであり，第二の課題は，独立話題分析によって独立な話題を抽出するためには，全てのデータを使用する必要があるというものである．

第 3 章

関連研究

本章では，独立話題分析の課題である 1) ユーザ制約を取り入れて話題を抽出する方法，および 2) 逐次増加するデータに対して適用できる話題抽出の方法についての関連研究についての紹介を行う．

3.1 ユーザ制約付き話題抽出

本節では，ユーザ制約を取り入れて話題抽出を行う方法やクラスタリングを行う方法について簡単に紹介する．

ユーザ制約には，制約付きクラスタリングの研究 [Basu08] で多く使われている，Must Link 制約と Cannot Link 制約が存在する．Must Link 制約とは，同じクラスタ（話題）に属すべきデータ（単語や文書）のペアに対し与えられるもので，Cannot Link 制約とは，別のクラスタ（話題）に属すべきデータ（単語や文書）のペアに対して与えられるものである．制約付きクラスタリングでは，これらの制約を全て，あるいは最大に満たすようにクラスタリングを行う．トピックモデルである LDA にこれらのユーザの制約を取り入れたものとして ITM (Interactive Topic Modeling)[Hu11] や LDA-DF (Latent Dirichlet Allocation with Dirichlet Forest priors)[Andrzejewski09] が提案されている．ITM では，LDA で文書，単語と話題の関係を示す確率モデルを求めた後に，単語間に Must Link 制約を加えたもので，話題の事前

分布をディリクレ分布ではなく，ディリクレ分布の一般化であるディリクレ木分布を用いて制約を実現している．LDA-DF では，あらかじめ単語間に与えられた Must Link 制約や Cannot Link 制約をディリクレ木分布を複数用いた混合分布であるディリクレ森分布を用いて実現している．さらに LDA-DF よりも任意の倫理表現をユーザ制約に利用できるように拡張した方法 [Kobayashi12] など，トピックモデルにユーザ制約を取り入れる方法についての研究は非常に盛んに行われている．

また，トピックモデルとは別の方法でユーザの制約を取り入れる方法も研究されている．Bar-Hillel らの提案する RCA (Relevant Component Analysis)[Bar-Hillel03, Bar-Hillel05] では，Must Link 制約と良く似た概念として，chunklet と呼ぶものを提案している．chunklet とは同じクラスタとなるべきデータの集合を指定したもので，RCA では，chunklet として与えられた集合内の距離が小さくなるような変換を計算し，クラスタリングを行う方法である．この RCA を拡張したものとして Hoi らの DCA (Discriminative Component Analysis)[Hoi06] がある．これは，chunklet として与えられた集合内の距離が小さくなり，かつ異なる集合間の距離が大きくなるような変換を計算する方法である．また Chang らは，同じクラスタとなるデータ集合を制約として与え，その制約が与えられたデータ同士の距離が 0 となるような変換を行い，データのクラスタリングを行う方法である LLMA (Locally Linear Metric Adaptation)[Chang04] を提案している．

他には，クラスタリング手法の代表的なものの一つである Kmeans[MacQueen67] に Must Link 制約や Cannot Link 制約を取り入れる方法も研究されている．Wagstaff らが提案する COP-Kmeans[Wagstaff01] は，Kmeans によって得られたクラスタリング結果を見て，ユーザがデータ間に Must Link 制約や Cannot Link 制約を与える．そして，それらの制約を満たすように再度 Kmeans を行う方法である．他の例としては，Basu らが提案する Seeded-Kmeans[Basu02] は予め与えられた制約を Kmeans の初期値としてクラスタリングを行うものや，Kmeans の目的関数として，Hidden

Random Markov Field (HMRF) を用い，これに Must Link 制約と Cannot Link 制約を取り入れたものを使用する HMRF-Kmeans[Basu04] という方法も提案されている．しかし，独立性の高い話題を求める独立話題分析にユーザ制約を加えるという研究は行われていない．

3.2 データ追加に基づいた話題抽出

本節では，データが逐次的に増加していく場合での話題抽出を行う方法について簡単に紹介する．また，増加するデータに主成分分析 (PCA; Principal Component Analysis) や独立成分分析 (ICA; Independent Component Analysis) を適用した方法についても紹介する．

増加していくデータから話題を抽出するトピックモデルの研究として，いくつかの方法が提案されている [Banerjee07]. 例えば，Neal らが提案した Online vMF[Neal98] がある．vMF とはフォンミーゼス-フィッシャー分布 (von Mises-Fisher distribution) を用いた混合モデルである．フォンミーゼス-フィッシャー分布の混合モデルでは，文書は対応する tf-idf を L2 正規化した単位ベクトルとして表される．つまり，全ての文書は超球面上にあるように配置される [Banerjee07]．Neal らはこの vMF モデルのアルゴリズムの高速化を行い，Online vMF を提案した．また，Banerjee は一般的な vMF モデルの特殊なケースである球形カーネルアルゴリズムに焦点をあて，完全な Online vMF を提案した．vMF モデルに新しい文書が与えられた場合，モデルのパラメータは新しい文書に基づいて更新する必要がある．更新を行うにはいくつかの方法があるが，完全な Online vMF では，現在の文書が割り当てられている混合成分のパラメータの部分だけ更新する簡単な方法を選択する方法である．

他の例としては，DCM を拡張した EDCM という方法が存在する [Banerjee07]．DCM (Dirichlet Compound Multinomial) モデルは指数関数的な分布ではないため，単純な再帰的な更新は DCM モデルの混合には適切ではない [Elkan06, Madsen05]．EDCM モデルとはこの DCM モデルの

指数関数近似を行ったものである．しかし，キュムラント関数が正確に定義されていないと，EDCM は実際には指数関数族モデルにはならない [Banerjee05, Azoury01]．そのため EDCM モデルは，窓の更新によってオンライン化を行っている．EDCM モデルに新しい文書が与えられた場合，そのモデルの構成要素から最も可能性の高い構成要素にその文書を割り当てる確率を計算する．その文書を最も可能性の高い要素に割り当てたあと，EDCM は各要素のパラメータを更新する．EDCM の各要素のパラメータは，新しい推定パラメータと今までのパラメータ値とで，窓をスライドさせ，そのときの移動平均値を用いて更新される手法である [Banerjee07].

また，Song らは incremental LDA を提案している [Song05]．この incremental LDA では，batch LDA として最初は小さな窓に入ってくる文書データが処理される．そのときパラメータは MAP 推定を用いて初期化される．このパラメータは新しく入力される全ての文書データに基づいて更新される．これによって推定されるトピックの割り当ては，累積された値と現在の文書内に含まれる単語にのみ依存して行われる．他には，PLSA を拡張した Online PLSA[Bassiou14] や Incremental PLSA[Chien07] など，様々なトピックモデルが増加するデータに対して適用できる方法が提案されている．

トピックモデルだけでなく，主成分分析 (PCA) にも増加するデータに対して適用できるアルゴリズム (IPCA; Incremental Principal Component Analysis) がいくつか提案されている [Hertz91, Oja85, Sanger89]．Weng らは fast IPCA のアルゴリズムである Covariance-Free IPCA (CCIPCA) を提案している [Weng03]．主成分分析 [Sirovich87] は LSI[Deerwester90] とともにデータ分析ではよく知られた方法である．CCIPCA では，古いデータと新しいデータの保持率は固定ではなく，動的に決定するためにアムネジック平均法に基づいて決定される．さらに，独立成分分析 (ICA) についても増加するデータに適用できる Online ICA[Schraudolph99] や Recursive ICA[Akhtar12] が提案されている．これらの ICA の方法は情報量最大化

(Infomax) アルゴリズム [Bell97, Amari96] を使用する方法である．しかし，独立性に尖度最大化を用いる独立話題分析を増加するデータへの適用方法についての研究は行われていない．

3.3 本章まとめ

本章では，独立話題分析の二つの課題との関連研究について紹介した．始めに，得られた話題がユーザの求める話題と異なる場合に，得られた話題にユーザ制約を加えて，ユーザ制約を満たす関連研究についていくつか紹介した．制約付きクラスタリングで一般的に使用されている Must Link 制約と Cannot Link 制約について述べ，それらの制約を加えた話題抽出手法など様々な方法について紹介した．続いて，データが逐次増加する場合における話題抽出に関する他の研究について述べた．話題抽出手法のトピックモデルだけでなく，主成分分析や独立成分分析において増加するデータへの適用方法についてもいくつか紹介した．これらの課題は，話題抽出では多く扱われているが，独立話題分析にはこれらの課題に取り組む研究は行われていないことについても述べた．

第 4 章

ユーザ制約付き独立話題分析

独立話題分析によって得られた話題がユーザの求める話題と異なる場合に，得られた話題にユーザの制約を加えて，ユーザ制約を満たしてかつ独立性の高い話題を得る方法を提案する [西垣 16b].

4.1 新たな制約の定義

ユーザ制約には, 制約付きクラスタリングの研究 [Basu08] では Must Link 制約と Cannot Link 制約が多く用いられているが，これらの制約は我々が行いたい制約とは異なる．我々が制約として与えたいことは，2 個の話題を 1 個に統合して話題の数を 1 個減らしたい場合と，1 個の話題を 2 個の話題に分離して話題の数を 1 個増やしたい場合である．しかし，Must Link 制約や Cannot Link 制約では話題の数は変更しないため，Must Link 制約や Cannot Link 制約とは異なる制約を考える必要がある．そこで，我々はユーザ制約として，次の 2 種類の制約を新たに定義する．**“2 個の話題を統合して 1 個の話題にして，話題の数を 1 個減らす制約”** を **Merge Link 制約**と，**“1 個の話題を分離して 2 個の話題にして，話題の数を 1 個増やす制約”** を **Separate Link 制約**と定義する．以上の二つの制約を満たし，かつ独立性の高い話題を求める方法について提案する．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する．

表 4.1　新たに定義する制約と制約付きクラスタリングでの一般的な制約との主な違い

	提案する制約	一般的な制約
	Merge Link/Separate Link	Must Link/Cannot Link
話題の数	**減少させる/増加させる**	変化なし
制約を与える対象	**話題**	文書

4.2　制約付きクラスタリングでの制約との違い

新たに提案する Merge Link 制約および Separate Link 制約と，制約付きクラスタリングでの一般的な制約である Must Link 制約および Cannot Link 制約との違いを簡単に表 4.1 に示す．

表 4.1 にまとめたように, 提案する制約である Merge Link 制約と Separate Link 制約は制約を与える前後で話題の数を変更する．一方で制約付きクラスタリングでの一般的な制約では話題の数は変化しない．また，制約を与える対象も異なっている．提案する制約では話題に制約を与えるが，一般的な制約では文書に対して制約を与える．

4.3　Merge Link 制約付き独立話題分析

4.3.1　目的

Merge Link 制約の例を，LA Times の論文データに対して独立話題分析を行って得た話題の重要単語を示した表 2.1 を用いて説明する．表 2.1 を見ると，話題 1 は「Revenue」，話題 2 は「Soccer」，話題 3 は「Foreign」，話題 4 は「Entertainment」，話題 5 は「Affairs」，話題 6 は「Stock」，話題 7 は「Team, Player」の内容を表していると考えられる．この時ユーザが，話題 1 と話題 6 はどちらも金融の話題を表しているので，1 個の話題「Finance」として，合計 6 個の新たな話題を得たい場合を考える．このようにユーザがある 2 個の話題を，1 個の話題に統合したいと考える時，その 2 個の話

題に Merge Link 制約を与えて，それを満たしてかつ独立性の高い話題を得る方法を提案する．

4.3.2 方法

Merge Link 制約を満たしてかつ独立性の高い話題を求める方法について考える．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おり，各列のベクトル $\mathbf{v}_t = (v_{1,t}, v_{2,t}, \ldots, v_{m,t})^{\mathrm{T}}$ が話題 t に対応している．また各話題は独立しているため，選択した 2 個の話題の平均は，その 2 個の話題の特徴をあわせ持ったものと仮定できる．また，制約とは関係のない残りの $k-2$ 個の話題は，更新後の変化を少なくしたい．FPICA[Hyvärinen99] では話題を 1 個ずつ求めていくため，制約を満たす話題と関係のない話題から求めることで，更新後の変化が少なくなると考えた．そこで，選択した 2 個の話題の平均とのコサインの絶対値が小さいものから順番に更新する方法を提案する．絶対値を取る理由は，2 個の話題の平均が他の話題にどの程度影響を与えているかを考えるためである．この考えを元に，提案する Merge 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．

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，独立話題分析のステップ (3) において，$k+1$ 個の成分を抜き出し，ステップ (4) で，$(k+1) \times m$ の行列 $\mathbf{X}$ を得る．ステップ (5a) において，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. ユーザが統合したい話題を 2 個選択する．選択した話題をそれぞれ，話題 i ($i \in \{1, \ldots, k\}$) と話題 j ($j \in \{1, \ldots, k\}$) とする．ただし，$i \neq j$ である．
3. Merge Link 制約を満たす新しい話題を話題 ℓ とし，話題 ℓ を得るた

めの回転ベクトルの初期値を次の式で設定する．

$$\mathbf{b} = (\mathbf{r}_i + \mathbf{r}_j)/2, \quad \mathbf{b} = \mathbf{b}/\|\mathbf{b}\|$$

ここで $\mathbf{r}_i$，$\mathbf{r}_j$ はステップ (1) で得た回転行列 $\mathbf{R}$ の i 列目と j 列目を表す．

4. ステップ (1) で得た回転行列 $\mathbf{R}$ の i 列目と j 列目を抜き出した $(k+1)\times(k-1)$ の行列を独立話題分析のステップ (5b) での $\mathbf{e}_t$ の代わりとして，話題 ℓ とのコサインの絶対値が小さいものから順に，独立話分析のステップ (5) と同様に FPICA に基づいて $1 \le t \le k-2$ までの新しい回転行列 $\dot{\mathbf{R}}$ を求める．
5. $t = k-2$ まで求めたら，最後にステップ (3) で求めた $\mathbf{b}$ を次の回転行列化を行う．

$$\mathbf{b} = \mathbf{b} - \dot{\mathbf{R}}\dot{\mathbf{R}}^{\mathrm{T}}\mathbf{b}, \quad \mathbf{b} = \mathbf{b}/\|\mathbf{b}\|$$

6. $\mathbf{b}$ を $\dot{\mathbf{R}}$ の最後の列に代入し，その時の $\dot{\mathbf{R}}$ が Merge Link 制約を満たす独立な話題の回転行列である．
7. 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 (7) を複数回繰り返すことで，話題を 1 個ずつ減らしていくことが可能である．Merge Link 制約付き独立話題分析のイメージを図 4.1 に図示する．この方法で Merge Link 制約を満たしてかつ独立性の高い話題を求める．

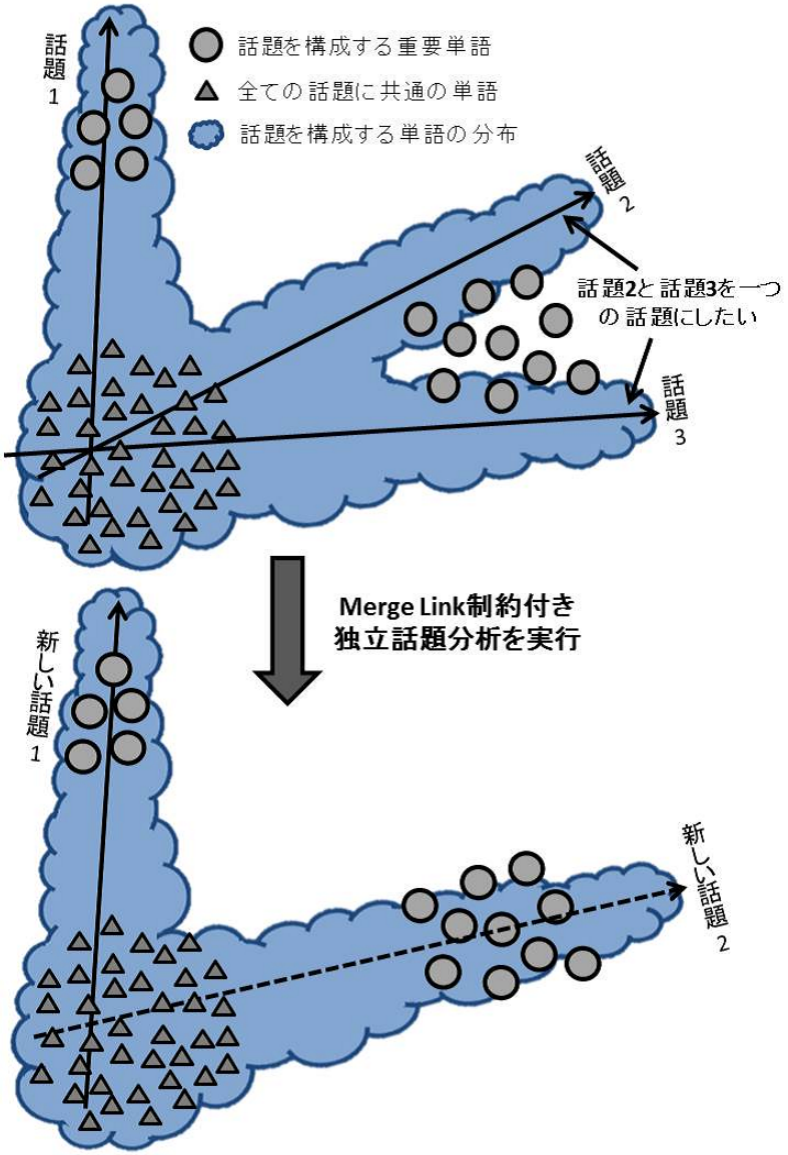


図 4.1　Merge Link 制約付き独立話題分析のイメージ

4.3.3 評価方法

前節で提案した，Merge Link 制約付き独立話題分析（以下，提案手法 (ML)）によって得られた話題が，制約を満たしているかどうかの検証および，得られた話題間の独立性についての評価のための実験を行う．Merge Link 制約の場合，話題間での類似度を測る尺度としてコサイン（詳細は後述）を用いる．話題間のコサインの絶対値が大きい場合，その話題間は近い内容を示している．提案手法 (ML) で得られた新しい 1 個の話題と，制約を与えた 2 個の話題とのコサインの絶対値が，他の話題とのコサインの絶対値よりも大きければ制約を満たしている．図 4.1 では，新しく得られた 1 個の話題（図 4.1 下の新しい話題 2）と制約を与えた 2 個の話題（図 4.1 上の話題 2 と話題 3）とのコサインの絶対値が，制約に関係のない他の話題（図 4.1 上の話題 1）とのコサインの絶対値よりも大きい．

Merge Link 制約を満たしているかどうかの評価には類似度としてコサインの絶対値を用いる．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 $\mathbf{v}_j = (v_{1,j}, \ldots, v_{m,j})^{\mathrm{T}}$ とのコサインは次のように定義される．

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

ここで $\mathbf{v}_j$ や $\mathbf{v}_i$ は，話題 j や話題 i における各単語の重要度を示すベクトルである．この各要素は，各単語が話題 j にどのような影響を与えているのかを示すものである．そのため，コサインの絶対値が大きければ，話題 i に影響を与えている単語と話題 j に影響を与えている単語は同じものが多いということになるので，話題 i と話題 j は近い話題であるとした．実験での評価では，Merge Link 制約によって新たに作成した話題 ℓ と，独立話題分析で得られた k 個の話題とのコサインを測る．そのとき，新たに作成した話題 ℓ と話題 i のコサインと，新たに作成した話題 ℓ と話題 j のコサインが他のものとのコサインよりも大きい場合，新たに作成した話題 ℓ は話

表 4.2　実験に適用する手法の主な違い

	話題の数		制約を満たす話題以外の話題の独立性最大化
	制約を入れる前	制約を入れた後	
既存手法	制約なしのため k		制約なし
比較手法 (ML)	$k+1$	k	**行わない**
提案手法 (ML)	$k+1$	k	**行う**

題 i と話題 j の特徴を持った話題であるとして，話題 ℓ は制約を満たしているとした．

また，得られた話題間の独立性について評価の方法には，相互情報量（詳細は後述）を用いて比較を行う．比較に使用する手法には，独立性を最大化する篠原の独立話題分析（以下，既存手法）と，制約を満たす話題の生成は提案手法と同様に行い，それ以外の話題は独立話題分析で得た話題を動かさない方法（以下，比較手法）を用いて行う．この方法は Merge Link 制約の場合，4.3.2 節のアルゴリズムで説明するとステップ (4) で回転行列の更新を行わず，ステップ (5) に進んで話題を求めたものが比較手法 (ML) となる．実験に適用する手法の主な違いを，表 4.2 にまとめる．

制約付き独立話題分析で得られる話題の独立性の評価については，得られた話題間の相互情報量を用いて比較を行う．話題間の相互情報量の値が 0 の場合，その話題間は完全に独立していることを意味し，相互情報量の値が小さい方が，話題間の独立性が高い．ただし，話題数が増えれば増えるほど，話題間が完全な独立でない限り，話題間の相互情報量は増加していく．得られた話題の単語の重要度の行列 ${}^*\mathbf{V}$ を用いて，[Brown12] の方法で求める．

実験には次のベンチマークデータを使用した．

- Los Angeles Times (LA Times) の新聞データ [Zhao02, Zhong03] で文書数は 6279，単語数は 31472 の文書データ

- DAILY KOS Blog (KOS Blog) のブログデータ [Lichman13] で文書数は 3430，単語数は 6906 の文書データ
- Neural Information Processing Systems (NIPS) の論文データ [Lichman13] で文書数は 1500，単語数は 12419 の文書データ

4.3.4 実験結果及び考察

4.3.2 節で提案した提案手法 (ML) で得られた話題が制約を満たしているかどうかの評価を行う．また，提案手法 (ML) で新たに得られた話題の独立性の評価も行う．

表 2.1 で示した話題 1 と話題 6 はどちらも金融の内容を表していると考えられるので，話題 1 と話題 6 を 1 個の話題となるように Merge Link 制約を与え，新しい話題を求めた時の各話題の重要単語を表 4.3 に示す．表 2.1 の話題 1 と話題 6 に Merge Link を与えて求めた話題である表 4.3 の話題 6 の重要単語が，表 2.1 の各話題の重要単語をどの程度含んでいるのかを表 4.4 に示す．表 4.4 は，表 4.3 の話題 6 を構成する重要単語は，表 2.1 の話題 1 を構成する重要単語の 60% と表 2.1 の話題 6 を構成する重要単語の 40% とで構成されていることを示している．そして，その時の両方に出現する重要単語を示している．このことから，表 4.3 の話題 6 の重要単語は表 2.1 の話題 1 と話題 6 の重要単語で構成されていることが分かる． また，表 2.1 での各話題と表 4.3 の話題 6 とのコサインの絶対値を表 4.5 に示す．表 4.5 の太字で示したように，表 4.3 の話題 6 は表 2.1 の話題 1 と話題 6 からのコサインの絶対値が大きいことが分かる．このことから，表 4.3 の話題 6 が制約を満たす話題であることが示せた．

次に，Merge Link 制約付き独立話題分析によって話題数 7 の時に Merge Link 制約を与えて 6 個の話題を得るものと，LA Times に話題数 6 として独立話題分析を行って得た話題時との違いについて考える．表 4.6 に，LA Times に話題数 6 として独立話題分析を適用して得られた話題の重要単語

表 4.3　表 2.1 の話題 1 と話題 6 に Merge Link 制約を与えて, **提案手法 (ML)** を適用して得られた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	rebound	league
2	aleen	macmin	art	entertain	report
3	soviet	afghanistan	israel	foreign	militari
4	polic	bush	counti	car	arrest
5	game	team	player	coach	bush
6	earn	million	quarter	stock	price

表 4.4　表 4.3 の話題 6 の重要単語と表 2.1 の各話題の重要単語の重なり

	表 4.3 の話題 6	両方に出現する重要単語
表 2.1 の話題 1	**0.6**	earn, million, quarter
表 2.1 の話題 2	0	N/A
表 2.1 の話題 3	0	N/A
表 2.1 の話題 4	0	N/A
表 2.1 の話題 5	0	N/A
表 2.1 の話題 6	**0.4**	stock, price
表 2.1 の話題 7	0	N/A

表 4.5　表 4.7 の話題 6 と表 2.1 の各話題とのコサインの絶対値

	表 4.3 の話題 6
表 2.1 の話題 1	**0.700**
表 2.1 の話題 2	0.001
表 2.1 の話題 3	0.056
表 2.1 の話題 4	0.031
表 2.1 の話題 5	0.020
表 2.1 の話題 6	**0.707**
表 2.1 の話題 7	0.060

表 4.6　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	tower
5	million	earn	bank	quarter	billion
6	polic	counti	offic	orang	citi

を示す．表 4.6 の話題 5 を見ると，表 4.3 の話題 6 のように，表 2.1 での話題 1 と話題 6 に与えた Merge Link 制約を満たす話題であることがわかる．この例の場合，話題数を 6 として独立話題分析を行うだけで Merge Link 制約を満たしていることがわかる．

ユーザが与える Merge Link 制約として，別の話題が選択される場合がある．例えば，表 2.1 の話題 3 と話題 6 を 1 個の話題として求めたい場合を考える．これは先程の例とは異なり，全く異なる話題を示していると考えられる 2 個の話題を選択している．表 2.1 の話題 3 と話題 6 を 1 個の話題となるように Merge Link 制約を与え，新しい話題を求めた時の各話題の重要単語を表 4.7 に示す．表 4.7 の話題 6 の重要単語が，表 2.1 の各話題の重要単語をどの程度含んでいるのかを表 4.8 に示す．表 4.8 は，表 4.7 の話題 6 を構成する重要単語は，表 2.1 の話題 3 を構成する重要単語の 20% と表 2.1 の話題 6 を構成する重要単語の 80% とで構成されていることを示している．そして，その時の両方に出現する重要単語を示している．このことから，表 4.7 の話題 6 の重要単語は表 2.1 の話題 3 と話題 6 の重要単語で構成されていることが分かる． また，表 2.1 での各話題と表 4.7 の話題 6 とのコサインの絶対値を表 4.9 に示す．表 4.9 の太字で示したように，表 4.7 の話題 6 は表 2.1 の話題 3 と話題 6 からのコサインの絶対値が大きいことが分かる．このことから，表 4.7 の話題 6 が制約を満たす話題であるとい

表 4.7　表 2.1 の話題 3 と話題 6 に Merge Link 制約を与えて, **提案手法 (ML)** を適用して得られた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	million	earn	quarter	revenu	net
2	scor	game	lead	rebound	league
3	aleen	macmin	art	entertain	report
4	bush	polic	budget	reagan	presid
5	game	team	polic	player	coach
6	soviet	bank	stock	rate	price

表 4.8　表 4.7 の話題 6 の重要単語と表 2.1 の各話題の重要単語の重なり

	表 4.7 の話題 6	両方に出現する重要単語
表 2.1 の話題 1	0	N/A
表 2.1 の話題 2	0	N/A
表 2.1 の話題 3	**0.2**	soviet
表 2.1 の話題 4	0	N/A
表 2.1 の話題 5	0	N/A
表 2.1 の話題 6	**0.8**	bank, stock, rate, price
表 2.1 の話題 7	0	N/A

うことが示せた.

この例でも同様に，Merge Link 制約付き独立話題分析によって話題数 7 の時に Merge Link 制約を与えて 6 個の話題を得たものと，LA Times に話題数 6 として独立話題分析を行って得た話題時との違いについて考える. 表 4.6 をみると，表 4.7 の話題 6 のように表 2.1 の話題 3 と話題 6 をあわせたような話題は存在しないことがわかる．この例では，先程の例とは異なり，話題数を 6 として独立話題分析を行うだけでは，Merge Link 制約を満たさないがわかる.

次に提案手法 (ML) で得られる話題の独立性について相互情報量を用い

表 4.9　表 4.7 の話題 6 と表 2.1 の各話題とのコサインの絶対値

	表 4.7 の話題 6
表 2.1 の話題 1	0.000
表 2.1 の話題 2	0.003
表 2.1 の話題 3	**0.676**
表 2.1 の話題 4	0.017
表 2.1 の話題 5	0.161
表 2.1 の話題 6	**0.705**
表 2.1 の話題 7	0.058

表 4.10　表 4.3 と表 4.7 時での各手法の相互情報量

	既存手法	提案手法 (ML)	比較手法 (ML)
表 4.3 の時	1.408	**1.117**	1.362
表 4.7 の時	1.408	**0.873**	1.096

て検証する．提案手法 (ML) で得られる話題が制約を満たしてかつ独立性の高いものとなっているのかを検証するために，既存手法によって同数の話題を抽出した場合と，比較手法 (ML) によって制約を満たす話題を抽出した場合とそれぞれ比較を行う．表 4.10 に話題数 6 での既存手法，提案手法 (ML)，比較手法 (ML) の三つの方法で求めた話題の相互情報量の値を示す．表 4.10 を見ると，太字で示した提案手法 (ML) によって得られた話題間の相互情報量は，既存手法や比較手法 (ML) によって得られた話題間の値よりも低い値となっていることから，独立性の高い話題が得られていることが分かる．

LA Times だけでなく他のベンチマークデータである KOS Blog と NIPS にも提案手法 (ML) を適用した．その時の相互情報量の値を表 4.11, 表 4.12 と表 4.13 にそれぞれ示す．既存研究である独立話題分析と提案した制約付き独立話題分析および比較手法は，ユーザの任意の数での独立な話題を抽出できる．また，Merge Link 制約は話題の数を減少させる制約であるた

め，同じベンチマークデータに対して話題の数を変更して各手法を適用した．最小の話題の数は 2 で，最大の数は適用するデータの文書数まで増やすことが可能である．

Merge Link 制約はユーザが各話題の重要単語を確認して制約を与える．そのため求める話題の数が増えるとユーザが確認する話題の重要単語の数も増えるため，制約を与えるのが非常に困難になる．そこで，本稿では話題数は 3 から 12 までの 10 個の結果のみを示す．また話題の数が 3 以上の理由は，Merge Link 制約を行った場合，最小の話題数 2 の時に制約を与えると，得られる話題数が 1 となってしまうためである．

既存手法と提案手法 (ML)，比較手法 (ML) のそれぞれを適用し，その時の相互情報量の値を表 4.11，表 4.12 と表 4.13 にそれぞれ示す．提案手法および比較手法においての制約は，Merge Link 制約の場合，制約を与える話題の組み合わせはランダムに選択し，話題数が増えた場合 10 回の平均値を示している．話題数は制約を与えて，最終的に得られた話題数を示している．つまり，話題数 5 だと，既存手法では話題数 5 で話題を求めている．提案手法 (ML) の場合，話題数 6 の時に制約を与えて，話題数 5 として話題を求めている．

表 4.11，表 4.12 と表 4.13 において，各ベンチマークデータで既存手法と提案手法 (ML)，比較手法 (ML) のそれぞれの方法によって得た話題の相互情報量の値を比較して，得られる話題の独立性を評価する．値を比較すると太字で示した提案手法 (ML) によって得られる話題間の相互情報量は，いずれのベンチマークデータにおいても比較手法 (ML) によって得られた話題間の相互情報量よりも低い値を示していることが分かる．したがって，各ベンチマークデータにおいて提案手法 (ML) によって得られる話題は，独立性の高いものとなっている．また，表 4.11，表 4.12 と表 4.13 には示していないが各話題数において，提案手法 (ML) によって得た話題のコサインの絶対値を比較すると，全て Merge Link 制約を満たしていることが確認できた．

表 4.11　LA Times に既存手法と提案する Merge Link 制約付き独立話題分析を適用した話題の相互情報量

	LA Times		
話題数	既存手法	提案手法 (ML)	比較手法 (ML)
3	0.238	**0.197**	0.236
4	0.497	**0.541**	0.631
5	0.853	**0.704**	0.706
6	1.127	**1.098**	1.163
7	1.408	**1.325**	1.412
8	1.804	**1.659**	1.789
9	2.437	**2.286**	2.345
10	2.963	**2.644**	2.925
11	3.453	**3.323**	3.429
12	4.030	**3.956**	4.024
⋮	⋮	⋮	⋮

表 4.12　KOS Blog に既存手法と提案する Merge Link 制約付き独立話題分析を適用した話題の相互情報量

	KOS Blog		
話題数	既存手法	提案手法 (ML)	比較手法 (ML)
3	0.189	**0.172**	0.2086
4	0.363	**0.292**	0.3489
5	0.679	**0.670**	0.6948
6	0.904	**0.892**	0.9054
7	1.229	**1.064**	1.0753
8	1.442	**1.437**	1.4754
9	1.856	**1.848**	1.8608
10	2.442	**2.416**	2.5332
11	3.248	**2.964**	3.1327
12	3.945	**3.598**	3.9633
⋮	⋮	⋮	⋮

表 4.13　NIPS に既存手法と提案する Merge Link 制約付き独立話題分析を適用した話題の相互情報量

	NIPS		
話題数	既存手法	提案手法 (ML)	比較手法 (ML)
3	0.237	**0.207**	0.211
4	0.367	**0.362**	0.400
5	0.615	**0.569**	0.592
6	0.854	**0.831**	0.843
7	1.164	**1.135**	1.189
8	1.439	**1.396**	1.463
9	1.802	**1.754**	1.818
10	2.237	**2.163**	2.173
11	2.738	**2.633**	2.731
12	3.319	**3.244**	3.258
⋮	⋮	⋮	⋮

表 4.14　既存手法，提案手法 (ML)，および比較手法の各種法によって得られる話題の違い

抽出した話題	既存手法	提案手法 (ML)	比較手法 (ML)
Merge Link 制約を満たす	×	○	○
高い独立性を持つ	○	○	×

以上により，提案手法 (ML) によって得られる話題は，ユーザ制約を満たし，かつ独立性の高い話題であることが示せた．また，既存手法によって話題数を変更し，提案手法 (ML) と同数の話題を得る方法は，独立性の高い話題を得ることができるが，ユーザ制約を満たす話題が得ることができるとは限らない．さらに，比較手法 (ML) によって得られる話題はユーザ制約を満たす話題であるが，独立性が高い話題であるとは言えないことがわかる．これらのことを表 4.14 にまとめる．

4.4 Separate Link 制約付き独立話題分析

4.4.1 目的

Separate Link 制約の例を，LA Times の論文データに対して独立話題分析を行って得た話題の重要単語を示した表 2.1 を用いて説明する．表 2.1 を見ると，話題 1 は「Revenue」，話題 2 は「Soccer」，話題 3 は「Foreign」，話題 4 は「Entertainment」，話題 5 は「Affairs」，話題 6 は「Stock」，話題 7 は「Team, Player」の内容を表していると考えられる．この時ユーザが，話題 5 を「Los Angeles」の話題と「President」の話題の 2 個に分離して，合計 8 個の話題を得たい場合を考える．このようにユーザがある 1 個の話題を，2 個の話題に分離したいと考える時，その 1 個の話題に Separate Link 制約を与えて，それを満たしてかつ独立性の高い話題を得る方法を提案する．

4.4.2 方法

Separate Link 制約を満たしてかつ独立性の高い話題を求める方法について考える．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．そして，選択された 2 個の重要単語となる独立な話題を 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，Merge Link 制約付き独立話題分析と同様に，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．

提案する 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．

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

分析のステップ (3) において，$k+1$ 個の成分を抜き出し，ステップ (4) で，$(k+1)\times m$ の行列 $\mathbf{X}$ を得る．ステップ (5a) において，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 ($z \in \{1,\ldots,k\}$) での行列 ${}^*\mathbf{V}$ の各要素の絶対値が 0 でない単語 p ($p \in \{1,\ldots,m\}$) と単語 q ($q \in \{1,\ldots,m\}$) をユーザが選択する．
3. 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．

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

$$\alpha = \arg\max_{1\le d\le n} a_{d,p}, \quad \alpha \in \{1,\ldots,n\}$$

ここで $\mathbf{u}_\alpha^{\mathrm{T}}$ は，行列 $\mathbf{U}$ の α 行目のベクトルの転置を表す．また $\alpha = \arg\max_{1\le d\le n} a_{d,p}$ は，行列 $\mathbf{A}$ での単語 p を示す列ベクトル $\mathbf{a}_p$ 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．

4. 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 (3) で求めた $\mathbf{c}$ を代入する．
5. 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のステップ (5b) での $\mathbf{e}_t$ の代わりとして，独立話題分析のステップ (5) 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．
6. 最後にステップ (3) 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．

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

$$\beta = \arg\max_{1\le d\le n} a_{d,q}, \quad \beta \in \{1,\ldots,n\}$$

7. ステップ (6) で得た $\mathbf{g}$ を次の回転行列化を行う．

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

8. $\mathbf{g}$ 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．
9. 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 (9) 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．Separate Link 制約付き独立話題分析のイメージを図 4.2 に図示する．この方法で Separate Link 制約を満たして独立性の高い話題を求める．

4.4.3 評価方法

前節で提案した，Separate Link 制約付き独立話題分析（以下，提案手法 (SL)）によって得られた話題が，制約を満たしているかどうかの検証および，得られた話題間の独立性についての評価のための実験を行う．

Separate Link 制約を満たしているかの評価には，各話題での単語の重要度を示す行列 $\mathbf{V}$ を用いる．Separate Link 制約で得られた単語 p が重要単語となる話題を話題 x，単語 q が重要単語となる話題を話題 y とする．話題 x および話題 y の重要度 $v_{p,x}$ と $v_{q,y}$ が，次の条件を満たせば話題 x および話題 y は制約を満たしている．

$$\arg\max_{1\le t\le k+1} |v_{p,t}| = x, \qquad \arg\max_{1\le t\le k+1} |v_{q,t}| = y$$

つまり，$\mathbf{v}_p = (v_{p,1}, \ldots, v_{p,k+1})$ の各要素の絶対値が最も大きい時の話題番号（列番号）が x ならば，話題 x は単語 p を重要単語とした制約を満たす話題である．同様に $\mathbf{v}_q = (v_{q,1}, \ldots, v_{q,k+1})$ の各要素の絶対値が最も大きい時の話題番号（列番号）が y ならば，話題 y は単語 q を重要単語とした制約を満たす話題である．図 4.2 では，制約を与えた 1 個の話題（図 4.2 上の話題 2）の二つの単語 p と単語 q の各話題での重要度を比較する．$v_{p,x}$ の絶対値

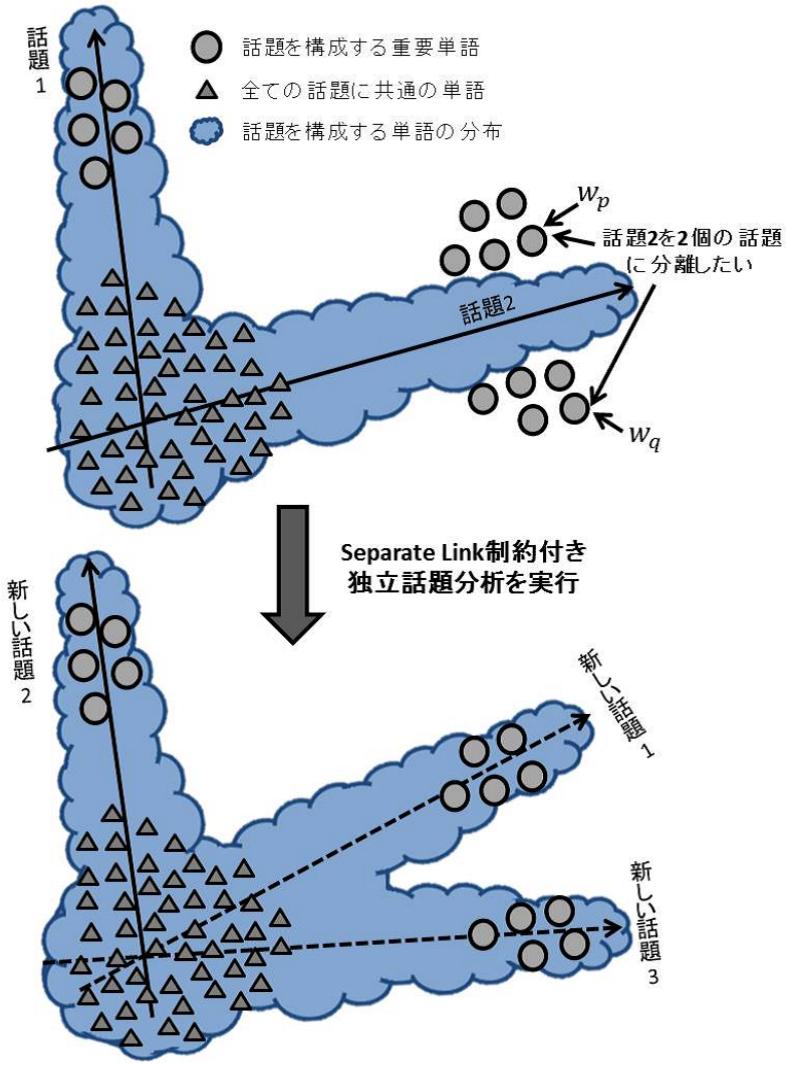


図 4.2　Separate Link 制約付き独立話題分析のイメージ

が話題 x の時（図 4.2 下の新しい話題 1）が最も大きく同様に $v_{q,y}$ の絶対値が話題 y の時（図 4.2 下の新しい話題 3）が最も大きいものとなっている．

また，得られた話題間の独立性について評価の方法には，Merge Link 制約付き独立話題分析の評価方法と同様に相互情報量を用いて比較を行う．比較に使用する手法には，独立性を最大化する篠原の独立話題分析（以下，既存手法）と，制約を満たす話題の生成は提案手法と同様に行い，それ以外の話題は独立話題分析で得た話題を動かさない方法（以下，比較手法）

表 4.15　実験に適用する手法の主な違い

	話題の数		制約を満たす話題以外の話題の独立性最大化
	制約を入れる前	制約を入れた後	
既存手法	制約なしのため k		制約なし
比較手法 (SL)	$k-1$	k	**行わない**
提案手法 (SL)	$k-1$	k	**行う**

を用いて行う．この方法は Separate Link 制約の場合，4.4.2 のアルゴリズムで説明するとステップ (5) で回転行列の更新を行わず，ステップ (6) に進んで話題を求めたものが比較手法 (SL) となる．実験に適用する手法の主な違いを，表 4.15 にまとめる．実験で使用するデータについても，Merge Link 制約付き独立話題分析の評価方法と同様である．

4.4.4　実験結果及び考察

4.4.2 で提案した提案手法 (SL) で得られた話題が制約を満たすかどうかの評価を行う．また，提案手法 (SL) で新たに得られた話題と既存手法で同数の話題を求めた時の話題間の独立性についての比較も行う．

表 2.1 で示した話題 5 の「Affairs」を更に 2 個の話題に分離して LA Times を分析したいと考えたので，話題 5 の arrest と presid に Separate Link 制約を与えた．その時の提案手法 (SL) で得られた各話題の重要単語を表 4.16 に示す．表 4.16 での話題 1 と話題 8 が表 2.1 での話題 5 に与えた制約を満たす「Los Angeles」の話題と「President」を示す話題である．また，表 4.16 での各話題での単語 arrest と presid の重要度を示す $|v_{\text{arrest},t}|$ と $|v_{\text{presid},t}|$ の値を表 4.17 に示す．表 4.17 の太字で示したように，arrest は話題 1 で値が最も大きく，presid は話題 8 で値が最も大きいことが分かる．このことから，表 4.16 の話題 1 と話題 8 が制約を満たす話題であることが示せた．

次に，Separate Link 制約付き独立話題分析によって話題数 7 の時に

表 4.16　表 2.1 の話題 5 に Separate Link 制約を与えて，**提案手法 (SL)** を適用して得られた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	pilic	bush	counti	car	arrest
2	million	earn	quarter	revenu	net
3	scor	game	lead	rebound	league
4	soviet	afghanistan	israel	foreign	militari
5	aleen	macmin	art	entertain	report
6	stock	price	bank	market	rate
7	game	team	player	coach	bowl
8	bush	counti	citi	presid	budget

表 4.17　表 4.16 での各話題での単語 polic と presid の重要度 $|v_{w,t}|$

	w = “arrest”	w = “presid”
$v_{w,\text{話題 1}}$	**65.502**	15.594
$v_{w,\text{話題 2}}$	0.633	3.105
$v_{w,\text{話題 3}}$	1.654	1.884
$v_{w,\text{話題 4}}$	9.965	5.284
$v_{w,\text{話題 5}}$	2.962	4.288
$v_{w,\text{話題 6}}$	18.544	5.754
$v_{w,\text{話題 7}}$	8.463	9.669
$v_{w,\text{話題 8}}$	12.345	**20.207**

Separate Link 制約を与えて 8 個の話題を得るものと，LA Times に話題数 8 として独立話題分析を行って得た話題時との違いについて考える．表 4.18 に，LA Times に話題数 8 として独立話題分析を適用して得られた 8 個の話題の重要単語を示す．表 4.18 の話題 1 と話題 6 を見ると，表 4.16 での話題 1 と話題 8 のように，表 2.1 での話題 5 に与えた Separate Link 制約を満たす話題であることがわかる．つまりこの例の場合，話題数を 8 として独立話

表 4.18　LA Times に独立話題分析を適用して得られた 8 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	bush	tower	senat	reagan	presid
2	million	earn	quarter	revenu	net
3	scor	game	lead	rebound	league
4	soviet	afghanistan	israel	foreign	afghan
5	aleen	macmin	art	entertain	report
6	polic	arrest	car	offic	kill
7	stock	bank	price	market	rate
8	game	team	bowl	player	coach

題分析を行うだけで Separate Link 制約を満たしていることがわかる．

ユーザが与える Separate Link 制約として，別の話題が選択される場合がある．例えば，話題 6 を 2 個の話題に分けたい場合を考える．これは先程の例とは異なり，ランダムな 1 個の話題からランダムに重要単語 2 個を選択し，2 個の話題に分ける場合を考えている．話題 6 の単語の compani と budget に Separate Link 制約を与えて，新しい話題を求めた時の各話題の重要単語を表 4.19 に示す．表 4.19 を見ると，話題 1 と話題 8 が表 2.1 での話題 6 に与えた制約を満たす 2 個の話題である．また，表 4.19 での各話題での単語 compani と budget の重要度を示す $|v_{\text{compani},t}|$ と $|v_{\text{budget},t}|$ の値を表 4.20 に示す．表 4.20 の太字で示したように，compani は話題 1 で値が最も大きく，budget は話題 8 で値が最も大きいことが分かる．このことから，表 4.19 の話題 1 と話題 8 が制約を満たす話題であることが示せた．

この例でも同様に，Separate Link 制約付き独立話題分析によって話題数 7 の時に Separate Link 制約を与えて 8 個の話題を得たものと，LA Times に話題数 8 として独立話題分析を行って得た話題時との違いについて考える．表 4.18 をみると，表 4.19 の話題 1 と話題 8 のように，表 2.1 の話題 6 を 2 個の話題に分離したような話題は存在しないことがわかる．この例で

表 4.19　表 2.1 に Separate Link 制約（話題 6）を加えた**提案手法 (SL)** を適用して得られた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	million	earn	quarter	rose	compani
2	earn	million	stock	price	rate
3	scor	game	lead	rebound	quarter
4	soviet	afghanistan	israel	foreign	counti
5	aleen	macmin	art	entertain	report
6	polic	bush	counti	car	arrest
7	game	bush	team	player	coach
8	bush	counti	presid	budget	citi

表 4.20　表 4.19 での各話題での単語 compani と budget の重要度 $|v_{w,t}|$

	w = “compani”	w = “budget”
$v_{w,}$ 話題 1	**24.823**	5.094
$v_{w,}$ 話題 2	14.325	3.086
$v_{w,}$ 話題 3	3.499	2.900
$v_{w,}$ 話題 4	0.875	0.583
$v_{w,}$ 話題 5	2.684	1.325
$v_{w,}$ 話題 6	8.558	18.174
$v_{w,}$ 話題 7	5.311	15.573
$v_{w,}$ 話題 8	11.946	**18.394**

は，先程の例とは異なり，話題数を 8 として独立話題分析を行うだけでは，Separate Link 制約を満たさないがわかる．

次に提案手法 (SL) で得られる話題の独立性について相互情報量を用いて検証する．Merge Link 制約の場合と同様に，提案手法 (SL) が制約を満たしてかつ独立性の高いものとなっているかを検証するために，既存手法によって同数の話題を抽出した場合と，比較手法 (SL) によって制約を満たす

表 4.21　表 4.16 と表 4.19 時での各手法の相互情報量

	既存手法	提案手法 (SL)	比較手法 (SL)
表 4.16 の時	1.804	**2.286**	2.320
表 4.19 の時	1.804	**2.386**	2.434

話題を抽出した場合とそれぞれ比較を行う．表 4.21 に話題数 8 での既存手法，提案手法 (SL)，比較手法 (SL) の三つの方法で求めた話題の相互情報量の値を示す．表 4.21 を見ると，太字で示した提案手法 (SL) によって得られた話題間の相互情報量は，既存手法によって得られた話題間の値よりも高いものの，比較手法 (SL) によって得られた話題間の値よりも低い値となっていることから，独立性の高い話題が得られていることを示した．

LA Times だけでなく他のベンチマークデータである KOS Blog と NIPS にも提案手法 (SL) を適用した．その時の相互情報量の値を表 4.22，表 4.23 と表 4.24 にそれぞれ示す．既存研究である独立話題分析と提案した制約付き独立話題分析および比較手法は，ユーザの任意の数での独立な話題を抽出できる．また，Separate Link 制約は話題の数を増加させる制約であるため，同じベンチマークデータに対して話題の数を変更して各手法を適用した．最小の話題の数は 2 で，最大の数は適用するデータの文書数まで増やすことが可能である．

Separate Link 制約はユーザが各話題の重要単語を確認して制約を与える．そのため求める話題の数が増えるとユーザが確認する話題の重要単語の数も増えるため，制約を与えるのが非常に困難になる．そこで，本稿では話題数は 3 から 12 までの 10 個の結果のみを示す．また話題の数が 3 以上の理由は，Separate Link 制約を行った場合，最小の話題数 2 に制約を与えると，得られる話題数が 3 となるためである．

既存手法と提案手法 (SL) と比較手法 (SL) のそれぞれを適用し，その時の相互情報量の値を表 4.22，表 4.23 と表 4.24 にそれぞれ示す．提案手法および比較手法においての制約は，Separate Link 制約の場合は，制約を与え

る話題はランダムに選択し，その話題中の単語は $^{*}\mathbf{V}$ の各要素の絶対値が上位 30 までの単語からランダムに 2 個選択する．これを 10 回繰り返した平均を示している．話題数は制約を与えて，最終的に得られた話題数を示している．つまり，話題数 5 だと，既存手法では話題数 5 で話題を求めている．提案手法 (SL) の場合，話題数 4 の時に制約を与えて，話題数 5 として話題を求めていることを示す．

表 4.22，表 4.23 と表 4.24 において，各ベンチマークデータで既存手法と提案手法 (SL)，比較手法 (SL) のそれぞれの方法によって得た話題の相互情報量の値を比較して，得られる話題の独立性を評価する．

値を比較すると太字で示した提案手法 (SL) によって得られる話題間の相互情報量は，いずれの場合においても比較手法 (SL) によって得られた話題間の相互情報量よりも低い値を示していることが分かる．したがって，各ベンチマークデータにおいて提案手法 (SL) によって得られる話題は，独立性の高いものとなっている．また，表 4.22，表 4.23 と表 4.24 には示していないが各話題数において，提案手法 (SL) によって制約として与えた各単語の各話題での重要度 $|v_{w,t}|$ を比較すると，全て Separate Link 制約を満たしていることが確認できた．

以上により，提案手法 (SL) によって得られる話題は，ユーザ制約を満たし，かつ独立性の高い話題であることが示せた．また，既存手法によって話題数を変更し，提案手法 (SL) と同数の話題を得る方法は，独立性の高い話題を得ることができるが，ユーザ制約を満たす話題が得ることができるとは限らない．さらに，比較手法 (SL) によって得られる話題はユーザ制約を満たす話題であるが，独立性が高い話題であるとは言えないことがわかる．これらのことを表 4.25 にまとめる．

表 4.22　LA Times に既存手法と提案する Separate Link 制約付き独立話題分析を適用した話題の相互情報量

	LA Times		
話題数	既存手法	提案手法 (SL)	比較手法 (SL)
3	0.238	**0.600**	0.643
4	0.497	**0.611**	0.612
5	0.853	**1.201**	1.215
6	1.127	**1.498**	1.650
7	1.408	**1.840**	1.853
8	1.804	**2.450**	2.497
9	2.437	**2.763**	2.864
10	2.963	**3.765**	3.869
11	3.453	**3.843**	3.991
12	4.030	**4.670**	4.698
⋮	⋮	⋮	⋮

表 4.23　KOS Blog に既存手法と提案する Separate Link 制約付き独立話題分析を適用した話題の相互情報量

	KOS Blog		
話題数	既存手法	提案手法 (SL)	比較手法 (SL)
3	0.189	**0.317**	0.332
4	0.363	**0.505**	0.598
5	0.679	**0.712**	0.795
6	0.904	**1.109**	1.168
7	1.229	**1.362**	1.497
8	1.442	**1.914**	2.048
9	1.856	**2.181**	2.183
10	2.442	**2.655**	2.726
11	3.248	**3.453**	3.468
12	3.945	**4.433**	4.500
⋮	⋮	⋮	⋮

表 4.24　NIPS に既存手法と提案する Separate Link 制約付き独立話題分析を適用した話題の相互情報量

	NIPS		
話題数	既存手法	提案手法 (SL)	比較手法 (SL)
3	0.237	**0.349**	0.371
4	0.367	**0.501**	0.557
5	0.615	**0.799**	0.846
6	0.854	**1.030**	1.039
7	1.164	**1.322**	1.329
8	1.439	**1.711**	1.738
9	1.802	**1.985**	2.015
10	2.237	**2.440**	2.449
11	2.738	**2.916**	2.985
12	3.319	**3.345**	3.379
⋮	⋮	⋮	⋮

表 4.25　既存手法，提案手法 (SL)，および比較手法の各種法によって得られる話題の違い

抽出した話題	既存手法	提案手法 (SL)	比較手法 (SL)
Separate Link 制約を満たす	×	○	○
高い独立性を持つ	○	○	×

表 4.26　既存手法，提案手法，および比較手法の各種法によって得られる話題の違い

抽出した話題	既存手法	提案手法	比較手法
ユーザ制約を満たす	×	○	○
高い独立性を持つ	○	○	×

4.5　本章まとめ

本章では，独立話題分析の課題の一つである，得られる話題がユーザの求める話題と異なる場合があるという問題を解決する方法について考察し，ユーザ制約付き独立話題分析を提案した．ここではユーザ制約として新しく二つ定義した．2 個の話題を 1 個の話題に統合する Merge Link 制約と，1 個の話題を 2 個の話題に分離する Separate Link 制約である．これらの制約を満たしてかつ，独立性の高い話題を得る方法を提案した．評価実験において，提案手法したユーザ制約付き独立話題分析で得られる話題は，表 4.26 に示すように，ユーザ制約を満たして，かつ既存手法である独立話題分析で得られる話題と同様に高い独立性を持つ話題であることが示された．

第 5 章

データ追加に基づく独立話題分析

5.1　目的

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，増加したデータのみの独立性に基づいて更新することで，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を提案する [西垣 16a, Nishigaki17].

5.2　方法

本節では，逐次増加していくデータに対して適用することが可能にする，データ追加に基づく独立話題分析の方法について考える．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．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す．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．データが追加され増加したデータと抽出

した独立性の高い話題を合わせる，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．

提案する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．
   - 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 5.1 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．

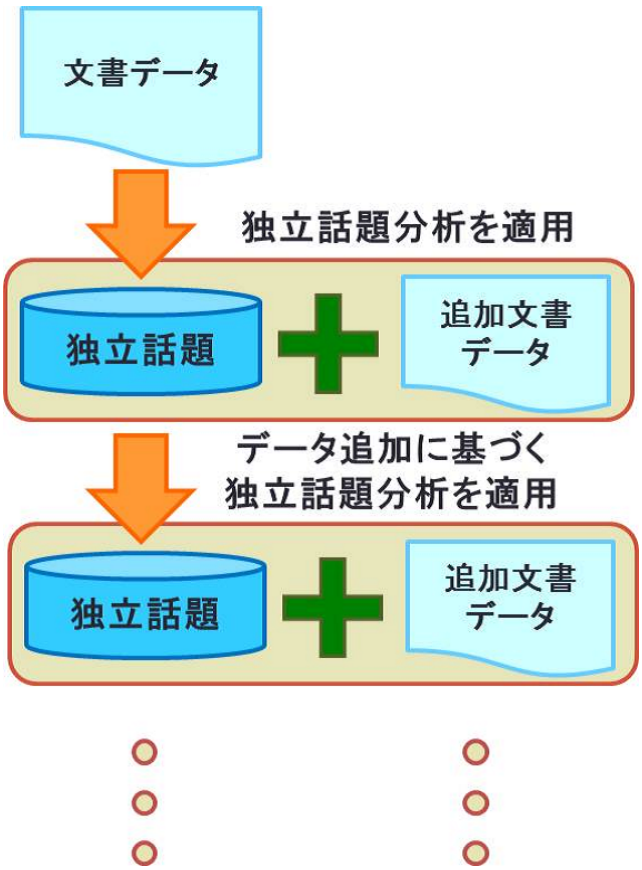


図 5.1　データ追加に基づく独立話題分析のイメージ

表 5.1　実験で使用した初期データの数

初期データ割合	50%	60%	70%	80%	90%
LA Times	3139	3767	4395	5023	5651
KOS Blog	1715	2058	2401	2744	3087

5.3　評価方法

前節で提案した，データ追加に基づく独立話題分析によって得られた話題が，全てのデータを用いて得られた話題にどれだけ近づけることが出来たかの検証および，その評価のための実験を行う．

実験には次のベンチマークデータを使用した．話題数が多いデータの Los Angeles Times の新聞データと話題数が少ないデータの KOS Blog のブログデータを用いる．

- Los Angeles Times (LA Times) の新聞データ [Zhao02, Zhong03] で文書数は 6279，単語数は 31472 の文書データ
- DAILY KOS Blog (KOS Blog) のブログデータ [Lichman13] で文書数は 3430，単語数は 6906 の文書データ

データ追加に基づく独立話題分析では，増加していくデータへの適用を考えているため，これらのベンチマークデータからランダムに 50%，60%，70%，80%，90% のデータをそれぞれ初期データとした．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．表 5.1 に使用したベンチマークデータの初期データ数を示す．

また，実験でユーザが求める話題数は，LA Times は 6 個，KOS Blog は 2 個とした．LA Times は，新聞データであるため多くの話題が存在し，[Zhao02, Zhong03] では話題数は 6 とあるので，話題数は 6 と設定した．一方で，KOS Blog データは政治に関することのみのデータであるため，話題

数は LA Times と比較すると少ない数の 2 と設定した．

データ追加に基づく独立話題分析によって得られた話題と全てのデータを用いて得られた話題との類似度を測る尺度としてコサインの絶対値を用いる．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 $\mathbf{v}_j = (v_{1,j}, \ldots, v_{m,j})^{\mathrm{T}}$ とのコサインは次のように定義される．

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

ここで $\mathbf{v}_j$ や $\mathbf{v}_i$ は，話題 j や話題 i における各単語の重要度を示すベクトルである．この各要素は，各単語が話題 j にどのような影響を与えているのかを示すものである．そのため，コサインの絶対値が大きければ，話題 i に影響を与えている単語と話題 j に影響を与えている単語は同じものが多いということになるので，話題 i と話題 j は近い話題であるとした．

実験での評価では，データ追加に基づく独立話題分析によって得られた話題と，全てのデータに対して独立話題分析を行って得られた話題とのコサインを測ることで行う．そのとき，データが増加するたびに，この値が大きくなっていれば提案するデータ追加に基づく独立話題分析によって得られた話題は，データ全てを用いて得られた話題に近づけることができたとした．

また，コサインの絶対値の総和は最大で話題数と同じ値となる．つまり，LA Times の場合は話題数が 6 なので最大値は 6 で，KOS Blog の場合は話題数が 2 なので最大値は 2 となる．この値が最大値の時，データ追加に基づく独立話題分析で求めた話題を構成する全単語の重みと，全てのデータを使用して求めた話題を構成する全単語の重みが全く同じということになる．

5.4　実験結果及び考察

5.2 節で提案したデータ追加に基づく独立話題分析で得られた話題が，全てのデータに独立話題分析を適用した話題にどれだけ近づけたのかの評価を行う．実験では，初期データとしてベンチマークデータの 50%，60%，70%, 80%, 90% をランダムで選択したものを使用する．また追加するデータ数は初期データ数よりも小さい数である 100 とする．

LA Times に提案手法を適用した結果を示す．図 5.2 に初期データ数を全体の 50% を使用したときの結果を，図 5.3 に初期データ数を全体の 60% を使用したときの結果を，図 5.4 に初期データ数を全体の 70% を使用したときの結果を，図 5.5 に初期データ数を全体の 80% を使用したときの結果を，および図 5.6 に初期データ数を全体の 90% を使用したときの結果をそれぞれ示す．

これらの図は，提案したデータ追加に基づく独立話題分析で得られた話題と全てのデータから抽出した独立性の高い話題とのコサインの絶対値の合計の 5 回の平均を示す．横軸はデータの追加回数，縦軸はコサインの絶対値の合計値である．コサインの絶対値の合計とは，提案手法で得られた話題 6 個と，全てのデータを用いて得られた話題 6 個のコサインの絶対値の合計となるので，最小値は 0，最大値は 6 となる．この値が 6 の場合，二つの手法で得られた話題を構成する全単語の重みはそれぞれまったく同じであるということである．

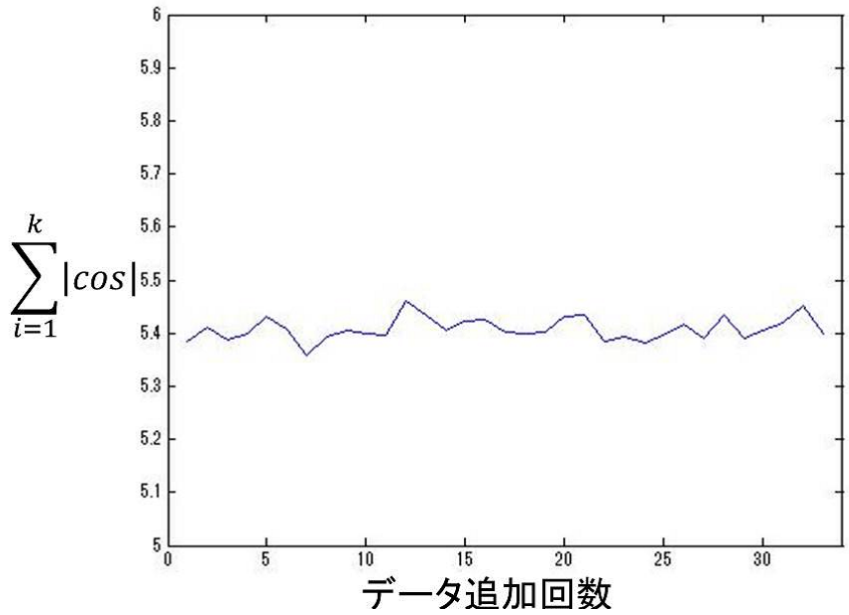


図 5.2　データ追加に基づく独立話題分析を LA Times に適用して抽出した話題と，全てのデータを使用して抽出した話題とのコサインの絶対値の総和．初期データとして LA Times のデータの 50% を使用し，残ったデータを追加データとして 100 ずつ追加していった．横軸はデータの追加回数，縦軸は全てのデータを使用して抽出した話題とのコサインの絶対値の総和であり，最大値は話題数が 6 であるため 6 となる．

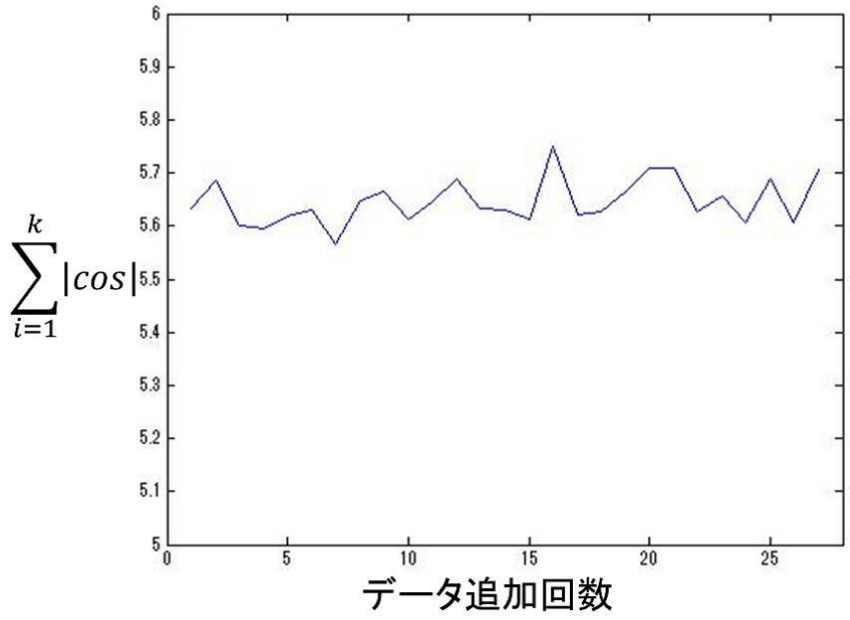


図 5.3　データ追加に基づく独立話題分析を LA Times に適用して抽出した話題と，全てのデータを使用して抽出した話題とのコサインの絶対値の総和．初期データとして LA Times のデータの 60% を使用し，残ったデータを追加データとして 100 ずつ追加していった．横軸はデータの追加回数，縦軸は全てのデータを使用して抽出した話題とのコサインの絶対値の総和であり，最大値は話題数が 6 であるため 6 となる．

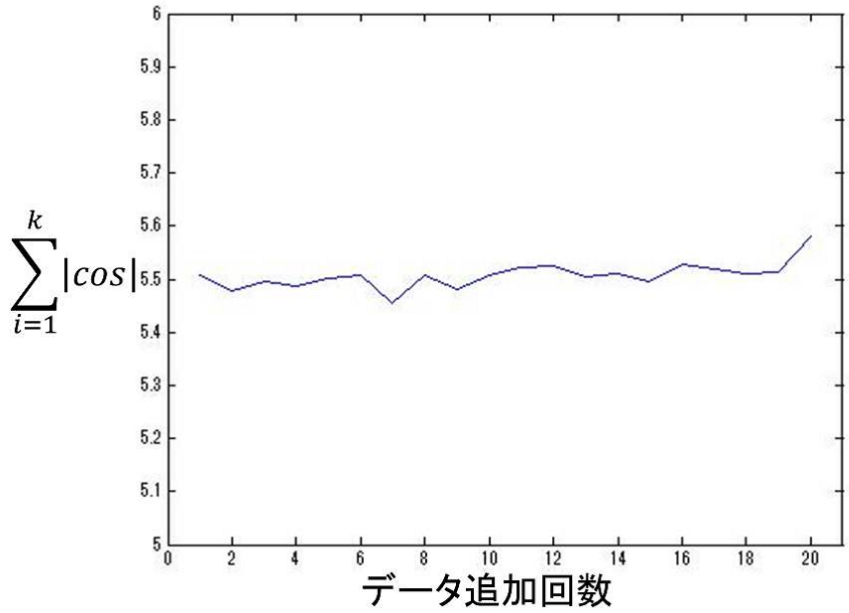


図 5.4　データ追加に基づく独立話題分析を LA Times に適用して抽出した話題と，全てのデータを使用して抽出した話題とのコサインの絶対値の総和．初期データとして LA Times のデータの 70% を使用し，残ったデータを追加データとして 100 ずつ追加していった．横軸はデータの追加回数，縦軸は全てのデータを使用して抽出した話題とのコサインの絶対値の総和であり，最大値は話題数が 6 であるため 6 となる．

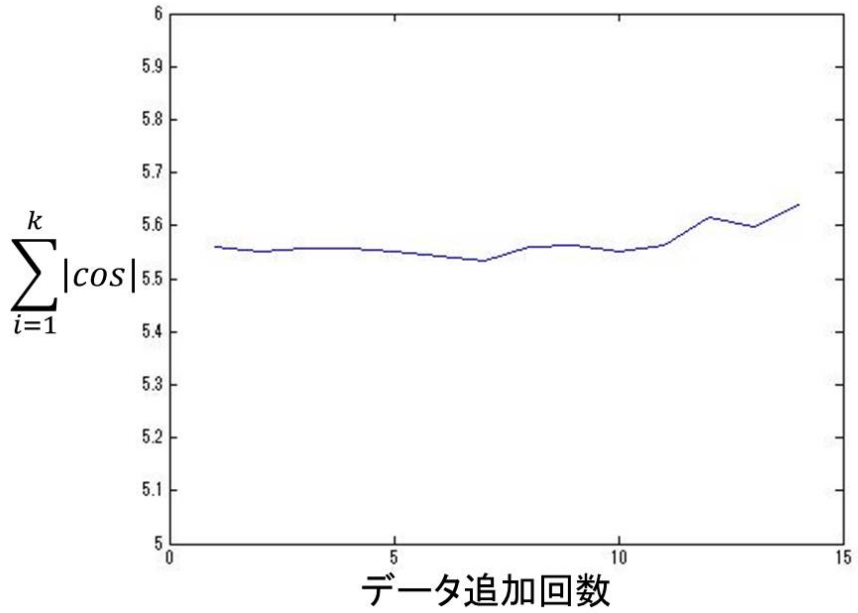


図 5.5　データ追加に基づく独立話題分析を LA Times に適用して抽出した話題と，全てのデータを使用して抽出した話題とのコサインの絶対値の総和．初期データとして LA Times のデータの 80% を使用し，残ったデータを追加データとして 100 ずつ追加していった．横軸はデータの追加回数，縦軸は全てのデータを使用して抽出した話題とのコサインの絶対値の総和であり，最大値は話題数が 6 であるため 6 となる．

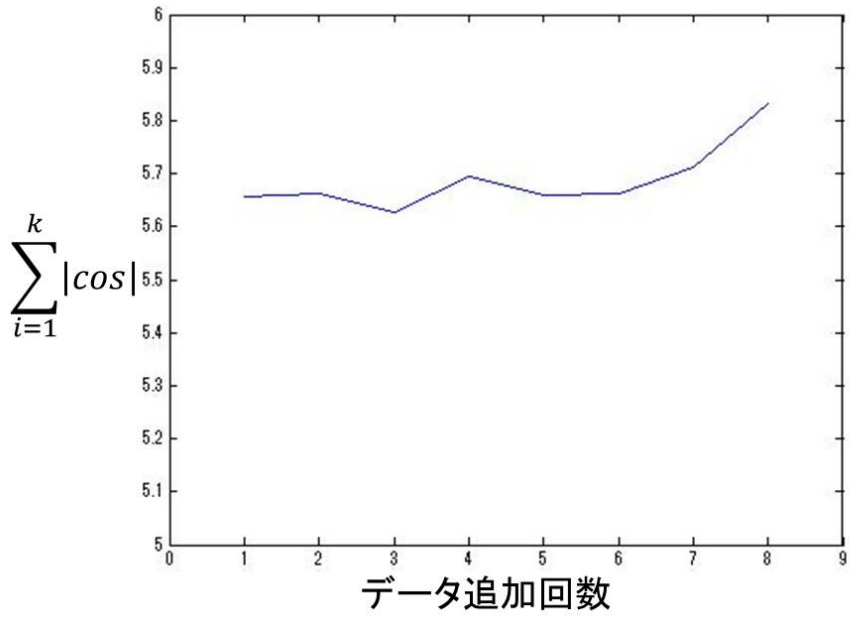


図 5.6　データ追加に基づく独立話題分析を LA Times に適用して抽出した話題と，全てのデータを使用して抽出した話題とのコサインの絶対値の総和．初期データとして LA Times のデータの 90% を使用し，残ったデータを追加データとして 100 ずつ追加していった．横軸はデータの追加回数，縦軸は全てのデータを使用して抽出した話題とのコサインの絶対値の総和であり，最大値は話題数が 6 であるため 6 となる．

表 5.2　LA Times 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 6 で適用して得られた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	league	rebound
4	soviet	afghanistan	israel	afghan	foreign
5	aleen	macmin	art	entertain	report
2	bank	loan	bush	stock	amp
3	earn	million	compani	quarter	revenu
6	polic	bush	arrest	diego	san

これらの図から，提案するデータ追加に基づく独立話題分析は初期データの数にかかわらず，データを追加する前（初期データを使用時）に抽出した独立な話題と全てのデータを使用して抽出した独立性の高い話題とのコサインの絶対値の総和が非常に高い値を示していることがわかる．これはつまり，それぞれの話題を構成する単語の重みの差異が小さいということである．

LA Times の場合では，全データの 50% を初期データとして使用する時（図 5.2）のコサインの値が，他の場合に比べると低い値を示し，その値は 5.38 である．この時の独立性の高い話題を構成する重要単語を表 5.2 に示す．また，話題数 6 として全データを使用した独立話題分析を適用したときの各話題の重要単語を表 5.3 に示す．表 5.2 を表 5.3 と比較すると，それぞれの話題がほとんど同じ話題を指し示していることがわかる．これらの話題を構成する重要単語を見ると，話題 1 は「Soccer」，話題 2 は「Foreign」，話題 3 は「Entertainment」，話題 4 は「National, Finance」，話題 5 は「Earn」，話題 6 は「Los Angeles」の内容を表していると考えられる．以上のことから，LA Times の場合，全データの 50% 程度使用することで，全データを使用したときとほとんど同じ話題を抽出することができると言える．

次に，提案したデータ追加に基づく独立話題分析によって抽出した話題

表 5.3　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	tower
5	million	earn	bank	quarter	billion
6	polic	counti	offic	orang	citi

が，データを追加し話題を更新していくことで，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ていくことができているかを確認する．図 5.2 から図 5.6 を見ると，いずれの場合においてもデータの追加回数を重ねる度に，徐々に大きな値となっていくことが確認できる．特に 70%, 80% および 90% の時のように初期データ数が多い時，その傾向が顕著であることがわかる．一方で，50% の時のように初期データ数が少ない時は，データの追加回数を増やしても値は安定して増加しているわけではない．これは 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．そのため初期データ数が少ないと，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．つまり，提案するデータ追加に基づく独立話題分析を適用する際は，初期データ内に全ての話題が含まれている必要がある手法であることがわかる．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は，LA Times の様に話題に属する文書数に差がある場合，初期データに使用するデータ数を増やしたほうが，話題を更新することで全てのデータを使用して抽出した独立性の高い話題に容易に近づけることができると考えられる．

LA Times のデータの時と同様に，次に KOS Blog に提案手法を適用した

結果を示す．図 5.7 に初期データ数を全体の 50% を使用したときの結果を，図 5.8 に初期データ数を全体の 60% を使用したときの結果を，図 5.9 に初期データ数を全体の 70% を使用したときの結果を，図 5.10 に初期データ数を全体の 80% を使用したときの結果を，および図 5.11 に初期データ数を全体の 90% を使用したときの結果をそれぞれ示す．

これらの図は，提案したデータ追加に基づく独立話題分析で得られた話題と全てのデータから抽出した独立性の高い話題とのコサインの絶対値の合計の 5 回の平均を示す．横軸はデータの追加回数，縦軸はコサインの絶対値の合計値である．コサインの絶対値の合計とは，提案手法で得られた話題 2 個と，全てのデータを用いて得られた話題 2 個のコサインの絶対値の合計となるので，最小値は 0，最大値は 2 となる．この値が 2 の場合，二つの手法で得られた話題を構成する全単語の重みはそれぞれまったく同じであるということである．

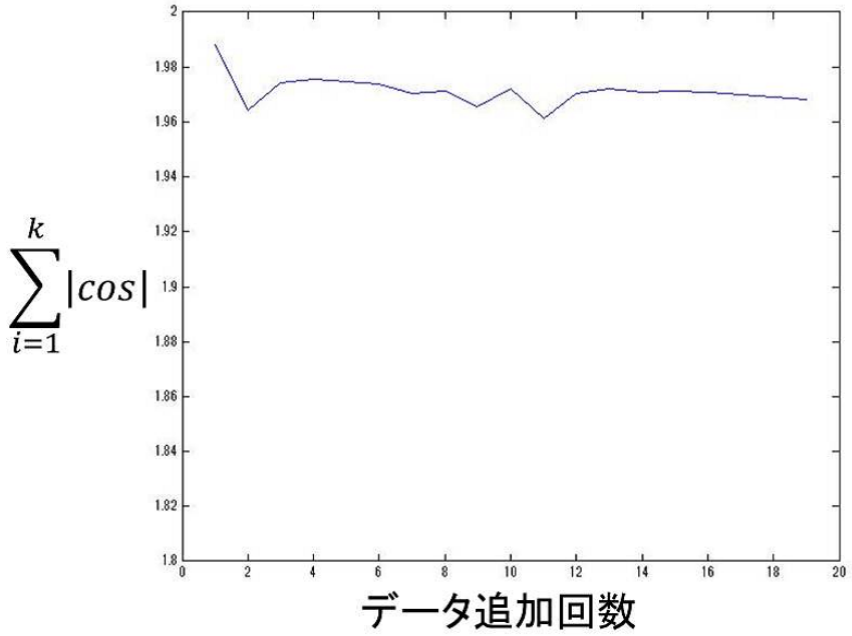


図 5.7　データ追加に基づく独立話題分析を KOS Blog に適用して抽出した話題と，全てのデータを使用して抽出した話題とのコサインの絶対値の総和．初期データとして KOS Blog のデータの 50% を使用し，残ったデータを追加データとして 100 ずつ追加していった．横軸はデータの追加回数，縦軸は全てのデータを使用して抽出した話題とのコサインの絶対値の総和であり，最大値は話題数が 2 であるため 2 となる．

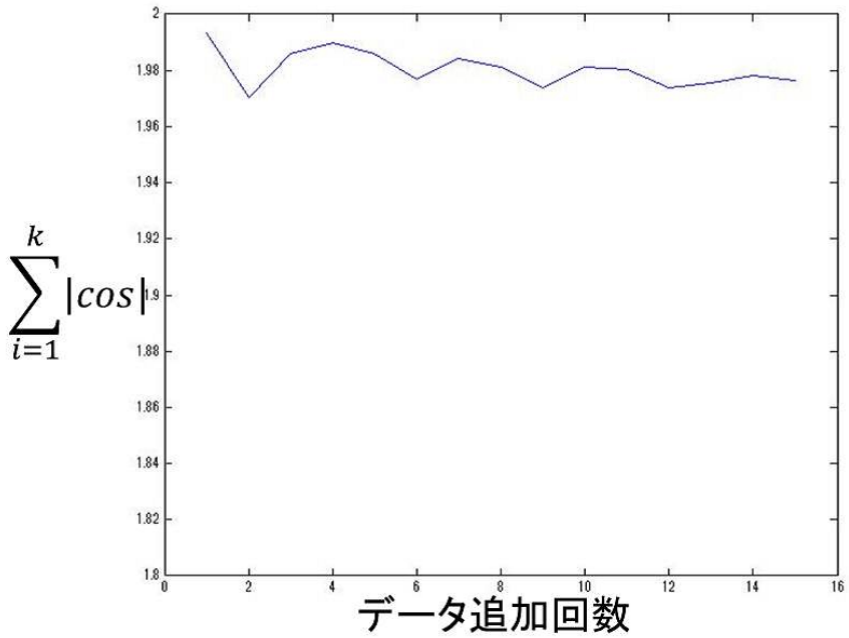


図 5.8　データ追加に基づく独立話題分析を KOS Blog に適用して抽出した話題と，全てのデータを使用して抽出した話題とのコサインの絶対値の総和．初期データとして KOS Blog のデータの 60% を使用し，残ったデータを追加データとして 100 ずつ追加していった．横軸はデータの追加回数，縦軸は全てのデータを使用して抽出した話題とのコサインの絶対値の総和であり，最大値は話題数が 2 であるため 2 となる．

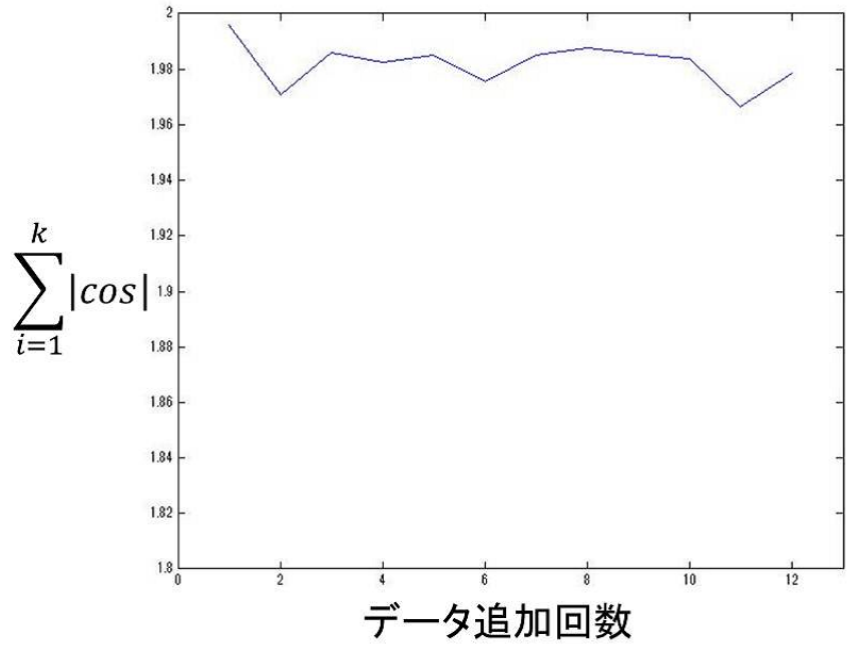


図 5.9　データ追加に基づく独立話題分析を KOS Blog に適用して抽出した話題と，全てのデータを使用して抽出した話題とのコサインの絶対値の総和．初期データとして KOS Blog のデータの 70% を使用し，残ったデータを追加データとして 100 ずつ追加していった．横軸はデータの追加回数，縦軸は全てのデータを使用して抽出した話題とのコサインの絶対値の総和であり，最大値は話題数が 2 であるため 2 となる．

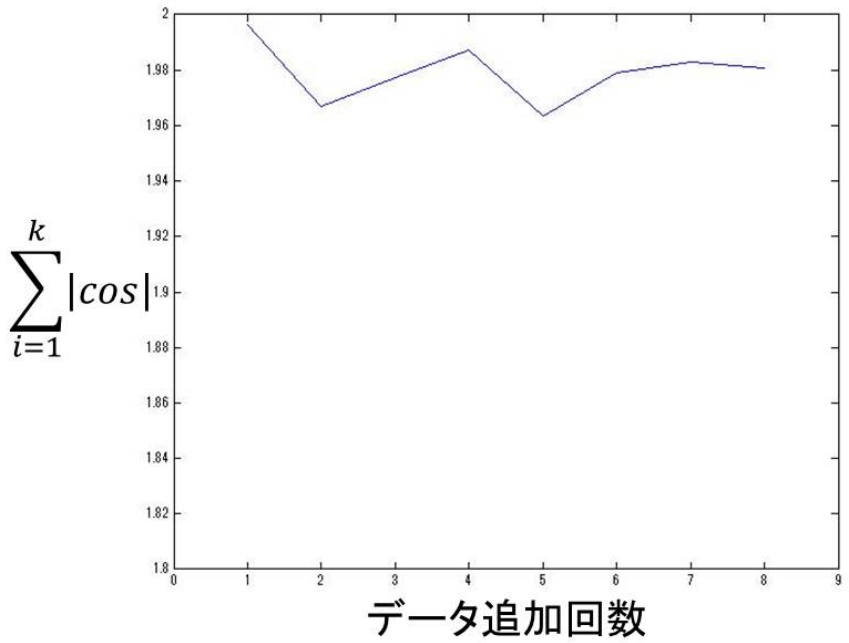


図 5.10　データ追加に基づく独立話題分析を KOS Blog に適用して抽出した話題と，全てのデータを使用して抽出した話題とのコサインの絶対値の総和．初期データとして KOS Blog のデータの 80% を使用し，残ったデータを追加データとして 100 ずつ追加していった．横軸はデータの追加回数，縦軸は全てのデータを使用して抽出した話題とのコサインの絶対値の総和であり，最大値は話題数が 2 であるため 2 となる．

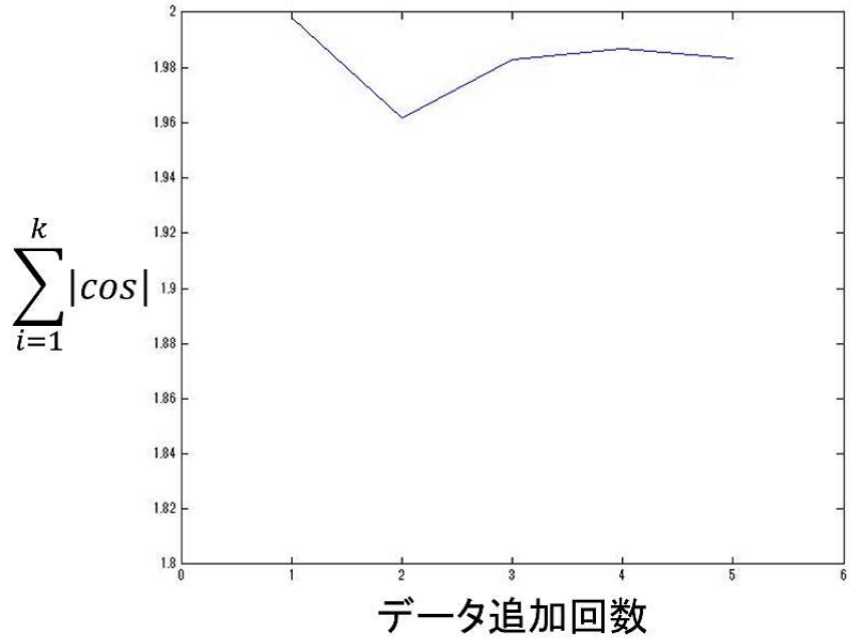


図 5.11　データ追加に基づく独立話題分析を KOS Blog に適用して抽出した話題と，全てのデータを使用して抽出した話題とのコサインの絶対値の総和．初期データとして KOS Blog のデータの 90% を使用し，残ったデータを追加データとして 100 ずつ追加していった．横軸はデータの追加回数，縦軸は全てのデータを使用して抽出した話題とのコサインの絶対値の総和であり，最大値は話題数が 2 であるため 2 となる．

表 5.4　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

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

表 5.5　KOS Blog 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

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

これらの図を見ると，KOS Blog の場合においても，LA Times の場合と同様に，提案するデータ追加に基づく独立話題分析は初期データの数にかかわらず，データを追加する前（初期データを使用時）に抽出した独立な話題と全てのデータを使用して抽出した独立性の高い話題とのコサインの絶対値の総和が非常に高い値を示していることがわかる．KOS Blog の場合では，全データの 50% を初期データとして使用した時のコサインの値は 1.96 であった．この時の独立性の高い話題を構成する重要単語を表 5.4 に示す．また，KOS Blog に対して話題数 2 として独立話題分析を適用したときの各話題の重要単語を表 5.5 に示す．表 5.4 を表 5.5 と比較すると，それぞれの話題がほとんど同じ話題を指し示していることがわかる．これらの話題を構成する重要単語を見ると，話題 1 は「Iraq War」，話題 2 は「Poll」の内容を表していると考えられる．以上のことから，KOS Blog の場合も，全データの 50% を使用するだけで，全データを使用したときと十分近い話題を抽出することができると言える．

次に，提案したデータ追加に基づく独立話題分析によって抽出した話題

が，データを追加し話題を更新していくことで，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ていくことができているかを確認する．図 5.7 から図 5.11 を見ると，いずれの場合においてもデータの追加回数を重ねても非常に高い値で推移していることが分かる．また，他より低い値を示した 1 回データが追加された時（横軸が 2 の時）より，いずれの場合においても，データを追加していくことで最終的に 0.02 程度値が大きくなっていることが確認できる．しかし，いずれの場合においても，初期データで求めた値が最も大きな値を示す結果となった．この原因として考えられることは，KOS Blog 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 5.4 および表 5.5 を見るとわかる．そして提案手法は，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少なからず受けてしまう．そのため，初期データで求めた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結果より，提案手法によって抽出した話題を，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．このことから，提案手法によって得られる話題を更新していくことで，増加するデータに適用することが困難であるという問題を解決する方法を提案した．しかし本提案手法は，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考えられる．

次に，独立話題分析を増加するデータへ適用する場合の計算量と，本提案手法の計算量を比較する．独立話題分析を増加するデータへ適用する場

合，初期データとして使用するデータ数を N，増加していくデータ数を M とすると，$O((N+M)^3)$ となる．一方で本提案手法の計算量は，追加するデータ L にのみ依存するので $O(L^3)$ となり，独立話題分析を増加するデータへ適用する場合よりも小さくなる．

本実験では，ベンチマークデータを使用しているため，独立話題の数は有限個でかつユーザが確認できる範囲の数で分析を行った．そのため，無限に文書データが存在する場合の独立話題の数については検討していない．感覚的には，無限に文書データが存在する場合でも，独立話題の数は無限に存在するわけではなく，数千程度で独立話題の数は頭打ちになるだろうと考えている．このように無限に文書データが存在する場合の独立話題の数を決定する研究や，推定する研究は行われていない．しかし本提案手法である，データ追加に基づく独立話題分析は，増加するデータへの適用を考えているため，無限に文書データが存在する場合における，独立な話題数の推定については今後検討していく必要がある．

5.5 本章まとめ

本章では，独立話題分析の課題の一つである，増加するデータに適用することが困難であるという問題を解決する方法について考察し，データ追加に基づく独立話題分析を提案した．初期データから抽出した独立話題を，データが増加する度に，増加したデータのみの独立性を用いて更新して，独立な話題を抽出する．評価実験では，データ追加に基づく独立話題分析によって抽出した話題と全てのデータを用いて抽出した独立話題の比較を行った．その結果，提案したデータ追加に基づく独立話題分析で得られた話題はデータが増加する毎に，全てのデータを用いて抽出した独立話題に近づけることができることが示された．またこの手法を適用する場合，話題数が多いデータのほうが望ましく，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を示した．

第 6 章

結論

本論文では，話題抽出手法の一つである独立話題分析に着目し，その独立話題分析によって得られる話題がユーザの望む話題と異なる場合が存在するという課題と，独立話題分析によって独立話題を得るためには全てのデータを使用する必要があるという課題の，二つの課題を解決する方法を提案した．

最初に，話題抽出全般について述べた．話題抽出は多くの研究がなされており，大きく分けて二つの種類の抽出方法が存在する．一つは確率的生成モデルであるトピックモデルである PLSA や LDA などが提案されており，もう一つは話題間の関係に着目して話題を抽出する LSI や独立話題分析が提案されている．様々な話題抽出の方法の中でも独立性に着目した方法として，トピックモデルである PLSA に話題と独立なセンシティブ情報という概念を取り入れた STI-PLSA や得られる話題間の独立性が高くなるように話題を抽出する独立話題分析が存在することを述べた．STI-PLSA は話題と独立なセンシティブ情報を求める方法で，独立性に着目した方法ではあるが，話題間の独立性については言及していない手法であった．本論文では，話題間の独立性について着目した独立話題分析，およびその課題について検討を行う，ということを述べた．

続いて，独立話題分析とそのアルゴリズムおよび課題について詳しく説明した．独立話題分析とは，膨大な量の文書データの中から信号処理の分

野の技術である独立成分分析を用いて，互いに独立な話題を求める手法である．しかしこの独立話題分析には二つ課題が存在する．第一の課題は，独立話題分析によって得られる話題は独立ではあるが，その話題がユーザの求める話題と異なる場合が存在するということである．第二の課題は，独立話題分析では独立な話題を求めるために，使用できる全てのデータを用いなければ独立な話題を得ることができないため，逐次増加するデータへの適用は難しいということである．本論文では，この二つの課題について解決する方法を提案することを述べた．

次に，独立話題分析が持つ二つの課題のそれぞれに関連する研究についてまとめた．はじめにユーザ制約付きの話題抽出についていくつかの方法を紹介し，制約付きクラスタリングで用いられる一般的な制約についてもここで述べた．その後，逐次増加するデータの中から話題を抽出する方法について紹介し，話題抽出の手法だけでなく，PCA や ICA などを増加するデータに適用する手法についてもここで簡単に説明した．しかし，これらの関連研究には独立話題分析にユーザ制約を加えるという方法や，データが逐次増加する場合での独立話題分析の適用方法については，研究が行われていないことについて指摘した．

続いて独立話題分析の二つの課題を解決する方法ついて提案した．はじめにユーザ制約付き独立話題分析として，新たに二つの制約を定義した．2 個の話題を 1 個の話題に統合する Merge Link 制約と 1 個の話題を 2 個の話題に分離する Separate Link 制約である．そして，それらの制約を満たす Merge Link 制約付き独立話題分析と Separate Link 制約付き独立話題分析を提案した．提案手法によって得られる話題は，それぞれユーザ制約を満たし，かつ独立性の高い話題であるということを，ベンチマークデータを用いた実験で示した．しかし，ユーザ制約である Merge Link 制約や Separate Link 制約をユーザが選択するため，求める話題の数が多くなるとユーザの負担が大きくなる．そのため，話題数に制限されないように，制約の与え方の改善や，グラフィカルユーザーインターフェース (GUI) について検討

する必要がある．

次に，逐次増加するデータに独立話題分析を適用できるデータ追加に基づく独立話題分析を提案した．この方法は，初期データに対して独立話題分析を行って得た独立話題を，データが追加される度に追加されたデータを用いて更新していくことで，全てのデータを使用して得られる独立な話題と同じ話題を求める方法である．評価のための実験として，テータ追加に基づく独立話題分析によって得られた話題と，全てのデータを使用して得られる話題の比較をベンチマークデータを用いて行った．実験の結果，提案するデータ追加に基づく独立話題分析によって得られる話題は，データの追加回数を重ねると徐々に全てのデータを使用して得られる話題に近づけることが出来ることを示した．また提案したデータ追加に基づく独立話題分析は，話題数が多く，また求めたい話題に属する文書データが初期データ内に含まれている必要があることを述べた．そのため，データが増加した時に新しい話題が増えるという場合に対処できない．そこでデータが増加した時に新たな話題が増加すればそれを検知し，求める話題数を変更できる方法について検討する必要がある．

本論文で対象にしている制約付き独立話題分析での独立性の高い話題の名称は，各話題での単語の重みを示す値が大きいものをユーザが確認して決定している．ユーザがそれぞれ単語を見て話題の名称を決定するため，話題数が多い場合のユーザの負担や，ユーザにどの程度単語を提示するのかなど，様々な課題が考えられる．そこで，単語の重みから自動的に話題の名称を決定する方法を検討する必要があるが，これについても課題が多いと考えている．単語の重みから自動的に話題の名称を決定する場合，重みの大きな単語の組み合わせから話題の名称を決定することになると考えられるが，その組み合わせと話題名称の対応の事前知識が必要になる．また，LA Times の新聞データのように政治，経済，スポーツなどの話題名称をユーザがイメージしやすいデータの場合，単語の重みから自動的に話題の名称を決定すると，その話題の名称がユーザが考える話題の名称と

同じ，あるいは比較的近いものになると考えられる．一方で，ユーザが単語から話題名称をイメージしにくいデータがある場合も考えられる．例えば，KOS Blog のようなデータなどのように，単一の話題で占められているデータから複数の独立性の高い話題を抽出した場合，事前知識に基づいて話題名称を自動的に決定するのは難しい．また仮に，何らかの方法で話題の名称を自動的に決定したとしても，その話題の名称とユーザが考える話題の名称とが大きく異なる場合が多いと考えられる．このように，ユーザが話題の名称をイメージしにくいデータに対しては，自動的に話題の名称を決定するのではなく，重みの大きい単語を並べたほうがユーザはその話題をイメージしやすいのではないかと考えている．

近年では，KOS Blog のようなユーザが話題の名称をイメージしにくいデータは，インターネット利用の増加に伴って爆発的に増えている．例えば Twitter や Facebook などの SNS 上でのデータやゲームや映画のレビューなどが挙げられる．こういったデータは誰でも書くことができ，また新聞などのように決められた話題があるわけではない．このようなデータから一般的な話題を求めて，それを提示しても，ユーザはそのデータを理解しにくいのではないかと考えられる．そこで，一般的な話題ではなく，独立性の高い話題をユーザに提示することで，ユーザはデータを理解しやすくなるのではないかと考えている．今後，制約付き独立話題分析によって抽出した独立性の高い話題が持つ印象についても検証を行いたいと考えている．

謝辞

指導教員である，東京工業大学総合理工学研究科知能システム科学専攻の新田克己教授には，研究に対する多大なる助言だけでなく，研究室生活においてもご支援いただきました．心より感謝申し上げます．今後ともどうぞよろしくお願いいたします．

外部研究指導教員である，青山学院大学理工学部経営システム工学科の小野田崇教授には，博士前期課程の頃から長きにわたり研究に対する多大なるご指導，ご支援いただきました．心より感謝申し上げます．今後ともどうぞよろしくお願いいたします．

新田研究室の岡田将吾助教には，研究についての助言や研究室生活においてご支援いただきました．心より感謝申し上げます．

新田研究室の中村敦子秘書，馬場裕子秘書には，研究室生活においてご支援いただきました．心よりお礼申し上げます．

新田研究室ならびに関連研究室の学生の皆さんには，研究に関する論議や研究室生活を通して楽しい時間を過ごさせていただきました．感謝いたします．今後ともよろしくお願いいたします．

最後になりますが，博士課程入学を快く承諾し，どのような状況においても応援してくれた父，母，妹に心から感謝いたします．

参考文献

[Akhtar12] MUhammad Tahir Akhtar, Tzyy-Ping Jung, Scott Makeig, and Gert Cauwenberghs. Recursive independent component analysis for online blind source separation. In *IEEE International Symposium on Circuits and Systems*, 2012.

[Amari96] Shunichi Amari, Andrzej Cichocki, and Howard Hua Yang. A new learning algorithm for blind signal separation. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *In Advances in Neural Information Processing Systems*, Vol. 8, pp. 757–763. The MIT Press, 1996.

[Andrzejewski09] David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 25–32. ACM, 2009.

[Azoury01] Katy S. Azoury and M. K. Warmuth. Relative loss bounds for online density estimation with the exponential family of distribution. *Machine Learning*, Vol. 42, No. 3, pp. 211–246, 2001.

[Banerjee05] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, Vol. 6, pp. 1705–1749, 2005.

[Banerjee07] Arindam Banerjee and Sugato Basu. Topic models over text streams: A study of batch and online unsupervised learning. In

SIAM International Conference on Data Mining, 2007.

[Bar-Hillel03] Aharon Bar-Hillel, Tomer Hertz, Noam Shental, and Daphna Weinshall. Learning distance functions using equivalence relations. In *Proceedings of the 20th International Conference on Machine Learning*, pp. 11–18, 2003.

[Bar-Hillel05] Aharon Bar-Hillel, Tomer Hertz, Noam Shental, and Daphna Weinshall. Learning a mahalanobis metric from equivalence constraints. *J. Mach. Learn. Res.*, Vol. 6, pp. 937–965, December 2005.

[Bassiou14] Nikoletta Bassiou and Constantine Kotropoulos. Online plsa: Batch updating techniques including out-of-vocabulary words. In *IEEE Transaction on Neural Networks and Learning Systems*, Vol. 25 of *11*, pp. 1953–1966, 2014.

[Basu02] Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. Semi-supervised clustering by seeding. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 27–34. Morgan Kaufmann Publishers Inc., 2002.

[Basu04] Sugato Basu, Mikhail Bilenko, and Mooney J. Raymond. A probabilistic framework for semi-supervised clustering. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 59–68. ACM, 2004.

[Basu08] Sugato Basu, Ian Davidson, and Kiri L. Wagstaff. *Constrained clustering: Advances in algorithms, theory, and application*. Chapman and Hall/CRC, 2008.

[Bell97] Anthony J. Bell and Terrence J. Sejnowski. An information-maximization approach to blind separation and blind dexonvolution. *Neural Computation*, Vol. 7, pp. 1129–1159, 1997.

[Blei03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent

dirichlet allocation. *The Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[Blei12] David M. Blei. Probabilistic topic models. *Commun. ACM*, Vol. 55, No. 4, pp. 77–84, 2012.

[Brown12] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luj*án*. Conditional likelihood maximisation: A unifying framework for information theoretic feature. *Journal of Machine Learning Research (JMLR)*, Vol. 13, pp. 27–66, 2012.

[Chang04] Hong Chang and DitYan Yeung. Locally linear metric adaptation for semi-supervised clustering. In *Proceedings of the 21st International Conference on Machine Learning*, pp. 153–160, 2004.

[Chien07] Jen-Tzung Chien and Meng-Sung Wu. Adaptive bayesian latent semantic analysis. In *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 16 of *1*, pp. 198–207, 2007.

[Deerwester90] Scott Deerwester, Susan T. Dumais, George W. Furnas, 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.

[Elkan06] Charles Elkan. Clustering documents with an exponential-family approximation of the dirichlet compund multinomial distribution. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[Hertz91] John A. Hertz, Anders S. Krogh, and Richard G. Palmer. *Introduction To the Theory of Neural Computation*. Addison-Wesley, 1991.

[Hofmann99a] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, UAI'99, pp. 289–296. Morgan Kaufmann Publishers Inc.,

1999.

[Hofmann99b] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22dd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pp. 50–57, 1999.

[Hoi06] Steven C.H. Hoi, Wei Liu, Michael R. Lyu, and WeiYing Ma. Learning distance metrics with contextual constraints for image retrieval. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2, pp. 2072–2078, 2006.

[Hu11] Yuening Hu, Jordan Boyd-Graber, and Brianna Satinoff. Interactive topic modeling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 248–257. Association for Computational Linguistics, 2011.

[Hyvärinen99] Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, Vol. 10, No. 3, 1999.

[Hyvärinen00] Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, Vol. 13, pp. 411–430, 2000.

[Hyvärinen01] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*. John Wiley & Sons, 2001.

[Kobayashi12] Hayato Kobayashi, Hiromi Wakaki, Tomohiro Yamasaki, and Masaru Suzuki. Topic models with logical constraints on words. In *Proceedings of Workshop on Robust Unsupervised and Semisupervised Methods in Natural Language Processing*, pp. 42–49, 2012.

[Lichman13] M. Lichman. Uci machine learning repository.

http://archive.ics.uci.edu/ml, 2013.

[MacQueen67] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pp. 281–297. University of California Press, 1967.

[Madsen05] Rasmus E. Madsen, David Kauchak, and Charles Elkan. Modeling word burstiness using the dirichlet distribution. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

[Neal98] Radford M. Neal and Geoffrey E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models*, pp. 355–368, 1998.

[Nishigaki17] Takahiro Nishigaki, Katsumi Nitta, and Takashi Onoda. Incremental learning of independent topic analysis. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, Vol. 11, No. 2, pp. 191–197, 2017.

[Oja85] Erkki Oja and Juha Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Application*, Vol. 106, pp. 69–84, 1985.

[Salton83] Gerard Salton, Edward A Fox, and Harry Wu. Extended boolean information retrieval. *Commun. ACM*, Vol. 26, No. 11, pp. 1022–1036, 1983.

[Sanger89] Terence D. Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural network. In *IEEE Transaction on Neural Networks*, Vol. 2, pp. 459–473, 1989.

[Schraudolph99] Nicol N. Schraudolph and Xavier Giannakopoulos. Online independent component analysis with local learning rate adaptation. In *NIPS*, pp. 789–795, 1999.

[Sirovich87] Lawrence Sirovich and Michael Kirby. Low-dimensional procedure for the caracterization of human faces. *Journal of Optical Society of America A*, Vol. 4, No. 3, pp. 519–524, 1987.

[Song05] Xiaodan Song, Ching-Yung Lin, Belle L. Tseng, and Ming-Ting Sun. Modeling and predicting personal information dissemination behavior. In *Proceedings of the 11th ACM SIGKKD International Conference on Knowledge Discovery and Data Mining*, 2005.

[Wagstaff01] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schroedl. Constrained k-means clustering with background knowledge. In *Proceedings of the 18th International Conference on Machine Learning*, pp. 577–584. Morgan Kaufmann, 2001.

[Weng03] Juyang Weng, Yilu Zhang, and Wey-Shiuan Hwang. Candid covariance-free incremental principal compoenent aalysis. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 25, No. 8, pp. 1034–1040, 2003.

[Zhao02] Ying Zhao and George Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *Conference of Information and Knowledge Management (CIKM)*, pp. 515–524. ACM, 2002.

[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 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. Technical Report OFS99-14, 信学技報, 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. Technical Report

R99036, 電力中央研究所報告, 2000.

[神嶌 06] 神嶌敏弘. 絶対クラスタリングと相対クラスタリング. 人工知能学会全国大会 (第 20 回) 論文集, No. 2A1-1, 2006.

[神嶌 15] 神嶌敏弘, 赤穂昭太郎, 佐藤一誠. 情報の独立性を強化したトピックモデル. 人工知能学会全国大会 (第 29 回) 論文集, No. 3L3-3, 2015.

[西垣 16a] 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, No. 2J3-5, 2016.

[西垣 16b] 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, Vol. 31, No. 4, pp. D–FB1 1–13, 2016.

[村田 05] 村田昇. 入門 独立成分分析. 東京電機大学出版局, 2005.

[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. Technical Report R02015, 電力中央研究所報告, 2003.

[渡部 03] 渡部勇. テキストマイニングの技術と応用 (特集 情報の分析・解析法). 情報の科学と技術, Vol. 53, No. 1, pp. 28–33, jan 2003.