The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

独立話題分析を用いた話題変遷の可視化及び話題追跡手法の提案

Visualization of topic transition and topic tracking method using independent topic analysis

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 理工学研究科
Aoyama Gakuin University Graduate School of Science and Engineering

In this paper, we propose a topic tracking method and visualization of topic transition using independent topic analysis. Independent topic analysis is a method for extracting mutually independent topics from the text data by using the independent component analysis. Documents continually delivered with times information such as news data are called time series documents. We often have a demand to grasp the topic transition and track the topic from large time series documents. But independent topic analysis is a method to analyze topics of one analysis data, and it is problem that not adapted to time series documents. So that , we propose a method to adapt independent topic analysis to time series documents. We visualized topic transitions by placing topics for each time series and establishing links based on similarity. In addition, we devised a topic tracking method that can track topics related to keywords. The topic tracking mainly keeps track of topic content and timing of topic appearance.

1. はじめに

インターネット上の情報提供, 配信サービスの進展により, 近年では, 電子ニュースなどネットワークを介したテキスト情報の配信が盛んに行われている. そのため, 膨大なテキスト情報が日々継続して得られ, 文書クラスタリングや情報抽出などのテキストマイニング技術が重要な研究課題となっている. 本稿では, テキストマイニングの課題の 1 つである, 話題分析に着目する. ここでの話題とは, 大量の文書間で複数の単語によって表現される情報である [佐藤 15]. 現在, 多くの話題分析手法が研究されているが, 本稿では, 話題間の関係性に着目した話題分析手法の 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] を用いる.

話題分析を用いることにより, 元の文書データを直接閲覧する場合と比較し, ユーザの労力を大幅に削減することが可能になる. しかし, 配信された大量の時系列文書の話題変遷の把握や話題追跡を行いたいというユーザの要求に対しては, 必ずしもこれらの手法が有効的であるとは限らない. 独立話題分析により, ある時刻, ある期間の文書データに含まれる話題は把握可能であっても, ユーザが話題変遷の把握や話題追跡を行うことは容易でない. 話題変遷の可視化及び話題追跡のためには, ある期間でどのような話題がみられるか, また時間の推移に伴い話題がどのように変化するかを可視化する手法が必要である.

2. 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 [西垣 16]. $\mathbf{V}$ 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，多くの単語や文書の重要度を 0 に近づけることができる．したがって, 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話題を求める．また独立話題分析で求める話題数 k はユーザ自身が与える．

3. 研究目的

本研究では電子ニュースのように, 時々刻々と配信される文書を時系列文書と呼ぶ。現在の独立話題分析は,1 つの文書データから独立性の高い話題を抽出する話題分析手法である. したがって, 電子ニュース等の時系列文書には適応していないことが課題の 1 つである. 独立話題分析により, ユーザは文書データに含まれる話題を把握することができる。しかし, 個々の文書データの話題は把握可能であっても, 時系列情報が表現されていないため, 話題変遷の可視化や話題追跡を行うことは容易でない. 話題変遷の可視化及び話題追跡のためには, 各期間に

連絡先: 山本健太，青山学院大学理工学研究科理工学専攻，
マネジメントテクノロジー,c5618178@aoyama.jp

おいてどのような話題がみられるか, また時間の変化に伴い話題がどのように変化するかを可視化する手法が必要である.

時系列文書に含まれる話題やトレンドの可視化技術及び, クラスタリング結果の分析技術などの関連研究は存在するが, 独立話題分析を用いた研究事例はない. 本稿では, 時系列文書のテキスト情報と時系列情報を基に話題変遷の可視化及び話題追跡が可能な手法を提案する.

4. 話題追跡手法

本稿は, 時系列文書の話題変遷の可視化及び話題追跡を行うための手法を提案する.

(1) 独立話題分析に基づく時系列文書の話題変遷の可視化手法の実装.
(2) 話題変遷及び詳細情報を可視化するための各種機能の実装
(3) 過去の話題との類似度判定による話題追跡手法の実装

4.1 流れ

時系列文書の話題変遷の可視化及び話題追跡の流れについて記述する.

(1) ユーザが設定した期間ごとの文書データに対し, 独立話題分析を行い, 各期間 k 個の話題を得る.
(2) 話題の詳細表示
(3) 各期間で得られた話題と隣接する 1 つ前の期間で得られた話題との類似性の有無を判定する.
(4) 話題追跡を行う場合には, 隣接する時刻間で類似性がない場合, 過去の時点で得られた話題との類似性を判定する.

4.2 話題間の類似性の判定

独立話題分析より各話題における内容を表す単語と各話題における単語の重要度から成るベクトルを作成する. ただし, 本稿では各話題における単語の重要度の絶対値の大きさが上位 N 個の単語だけを対象とする. 理由としては, 話題の内容と直接的に関係のない一般的な単語, 及びほとんど使われていない単語の影響を除くためである. これにより話題間の類似性がより明確に判定できるようになる.

隣接する期間の話題間の類似性をベクトルの類似性に基づいて判定する. 本稿では, コサイン類似度を用いた. 期間 t と期間 t-1 の話題のベクトルをそれぞれ $\mathbf{V}_i^{\mathrm{t}}(i \in (1,\ldots,k))$, $\mathbf{V}_j^{\mathrm{t-1}}(j \in (1,\ldots,k)$ とすると, この 2 つのベクトル間のコサイン類似度 $cos(\mathbf{V}_i^{\mathrm{t}}, \mathbf{V}_j^{\mathrm{t-1}})$ は次式で与えられる.

$$cos(\mathbf{V}_i^{\mathrm{t}}, \mathbf{V}_j^{\mathrm{t-1}}) = \frac{\mathbf{V}_i^{\mathrm{t}}\mathbf{V}_j^{\mathrm{t-1}}}{|\mathbf{V}_i^{\mathrm{t}}| \cdot |\mathbf{V}_j^{\mathrm{t-1}}|}$$

この値は, ベクトル間のなす角度のコサイン値であるため,$0 \leq cos(\mathbf{V}_i^{\mathrm{t}}, \mathbf{V}_j^{\mathrm{t-1}}) \leq 1$ となり, 値が 1 に近づくほど 2 つのベクトルは類似している. 隣接する期間 $\mathbf{V}_i^{\mathrm{t}}$,$\mathbf{V}_j^{\mathrm{t-1}}$ のすべての組み合わせに対してコサイン類似度を計算し, 閾値 S を超える場合に限り, 話題が類似していると判定する. 閾値 S はユーザが決定する. 期間 t の話題 i と期間 t-1 の話題 j が閾値以上の場合, 話題の類似といい, 式 (1) が成り立つ.

$$cos(\mathbf{V}_i^{\mathrm{t}}, \mathbf{V}_j^{\mathrm{t-1}}) > S \qquad i,j \in (1,\ldots,k) \tag{1}$$

4.3 話題の変化

話題追跡のために, 話題間の類似性の判定結果を用いて, 話題の類似に加えて,4 つの話題の変化を求める. 連続する期間の類似性を基に検出される話題の変化は, 話題の結合, 話題の分離がある. 連続しない期間の類似性も考慮して, 検出される話題の変化は, 話題の復活, 新規話題がある.

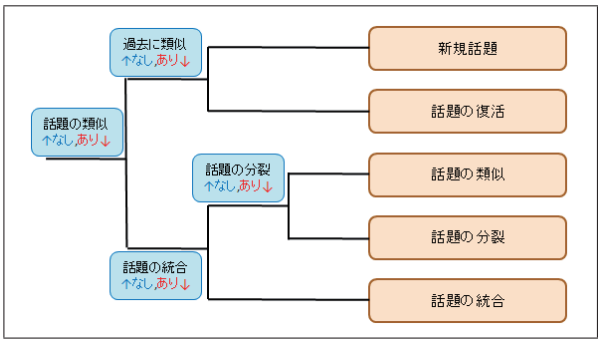

図 1: 話題の変化

4.3.1 話題の統合

話題の統合とは, 連続する期間 $(t, t-1)$ において期間 $t-1$ の 2 つ以上の話題が期間 t の話題 $\mathbf{V}_i^{\mathrm{t}}$ に話題の類似があることである. 話題の統合では以下の式 (2) が成り立つ. また話題の統合は話題数が減ることを意味していると考えられる.

$$cos(\mathbf{V}_i^{\mathrm{t}}, \mathbf{V}_j^{\mathrm{t-1}}) > S \quad \& \quad cos(\mathbf{V}_i^{\mathrm{t}}, \mathbf{V}_m^{\mathrm{t-1}}) > S \tag{2}$$

$$j, m \in (1,\ldots,k), j \neq m \tag{3}$$

4.3.2 話題の分裂

話題の分裂とは, 連続する期間 $(t, t-1)$ において, 期間 $t-1$ の話題 $\mathbf{V}_j^{\mathrm{t-1}}$ から, 期間 t の 2 つ以上の話題に対し, 話題の類似があることである. 話題の分裂では以下の式 (4) が成り立つ. また話題の分裂は話題数が増えることを意味していると考えられる.

$$cos(\mathbf{V}_i^{\mathrm{t}}, \mathbf{V}_j^{\mathrm{t-1}}) > S \quad \& \quad cos(\mathbf{V}_m^{\mathrm{t}}, \mathbf{V}_j^{\mathrm{t-1}}) > S \tag{4}$$

$$i, m \in (1,\ldots,k), i \neq m \tag{5}$$

4.3.3 話題の復活

話題の復活とは連続する期間 $(t, t-1)$ では類似する話題が存在せず, 過去の期間 $l, (l \in (1,\ldots,t-2))$ に話題の類似があることである. 話題の復活では以下の式 (6) が成り立つ話題の復活は, 季節関連やオリンピックなど定期的なイベント, 選挙などの不定期なイベントなどが抽出されると考えられる.

$$cos(\mathbf{V}_i^{\mathrm{t}}, \mathbf{V}_j^{\mathrm{t-1}}) < S \quad \& \quad cos(\mathbf{V}_i^{\mathrm{t}}, \mathbf{V}_m^{\mathrm{l}}) > S \tag{6}$$

$$l \in (1,\ldots,t-2), m \in (1,\ldots,k) \tag{7}$$

4.3.4 新規話題

新規話題とは連続する期間 $(t, t-1)$ では類似する話題が存在せず, 過去の期間 $l, (l \in (1,\ldots,t-2))$ にも類似する話題が存在しないことである. 新規話題では以下の式 (8) が成り立つ.

$$cos(\mathbf{V}_i^{\mathrm{t}}, \mathbf{V}_j^{\mathrm{t-1}}) < S \quad \& \quad cos(\mathbf{V}_i^{\mathrm{t}}, \mathbf{V}_j^{\mathrm{l}}) < S \tag{8}$$

$$l \in (1,\ldots,t-2), \tag{9}$$

4.4 話題変遷の可視化

話題変遷の可視化の流れについて記述する.

(1) 独立話題分析により得られた各期間の話題を, 時間軸上に配置する. その際, 各話題における単語の重要度の絶対値が最も高い単語をラベルとする.
(2) 連続する期間の話題の類似性を基に, 類似性が閾値以上の話題間にリンクを張る.
(3) リンクを張った話題間を同色にする, 類似性のない話題は白色とする.

表 1: 機能一覧

番号	内容
1	特定の単語に関する話題の追跡
2	特定の単語を含む記事の抽出
3	特定の単語の出現数の抽出
4	日付で話題結果の検索
5	文書検索
6	話題に影響を与える記事, 文書の抽出
7	話題の詳細表示 (重要単語の表示等)
8	記事本文, 見出しの表示

4.5 機能

時系列文書の話題変遷の可視化及び話題追跡を行う上で必要になる機能を実装する. 各機能単体での使用と複合しての使用により可視化及び話題追跡を行う.

5. 適用事例

5.1 データセット

毎日新聞社提供, 日外アソシエーツ発売の CD-毎日新聞データ集 [毎日新聞] の 2014 年から 2017 年の 4 年分のデータをデータセットに使用した. データセットの規模は, 記事数が 1 年あたり約 10 万件,1 日あたり約 150 件から 200 件である. データセットには記事 ID, 掲載面コード, 索引番号, 時系列情報や見出し, 記事のキーワード, 記事の本文などの情報が含まれている. 本稿では, 時系列情報と記事の本文から単語を抽出して話題分析に用いた.

5.2 閾値設定

本稿での話題の追跡では,1 日ごとの記事に対し独立話題分析を行い, 時系列ごとに配置した. また, 類似性の有無を判定するための閾値は 20 度として, 話題間のベクトル角度の差が 20 度以下の話題を類似と判定した. したがって閾値 $S = 0.94$ であり, 閾値より大きい類似度を持つ話題間には類似性があると判定した.

5.3 単語の抽出

形態素解析ソフト Janome を用いて形態素解析を行う. 日本語形態素解析して得られる単語から, 文書における単語の頻度行列 **A** を作成し, 独立話題分析に用いる. 独立話題分析の話題は分析データ **A** にどの単語を含めるかで抽出される話題が大きく変化する. 本稿では, 一般, 自立, 固有名詞等の名詞だけを対象とした. また, 標準の IPA 辞書では複合語はうまく処理できない. 名詞を連結して自動的に複合語とする手法があるが, ある名詞を含む様々な単語のバリエーションができることから, 辞書の追加を行った. 新聞データの分析に用いられる日経シソーラスを追加し, 形態素解析を行った.

5.3.1 分析対象外ワード

話題抽出において意味をなさない語や、話題抽出の精度を下げると考えられる単語を分析対象外ワードとして, 分析に用いる単語に制限を加えた. 一般的に新聞のデータは校正されているために,SNS データやレビューデータと比較して, ノイズの原因となる単語や言い回しは少ない. 分析対象外ワードは大きく 3 種類設定した. 1 つ目は数字を含む単語である. 数字を含む単語は”10 日” といった時系列データが含まれている場合や”9.78 秒” など記録や結果を表すことが多い. 結果や時系列データは話題の内容に影響を与えない単語のため, 分析対象から外す. 2 つ目はひらがなのみで構成されている単語である. Janome を用いた形態素解析では,” の” や” き” などひらがなが名詞として検出されることがある. また漢字の送り仮名も同様に抽出されているため, 特定の単語の出現率が 2 倍に増加する. 出現率の増加により話題抽出に影響を与えるため, ひらがなだけで構成されている単語を分析対象から外す. 3 つ目は 1 文字の単語である. 漢字 1 文字の単語や略語を分析対象から外す.

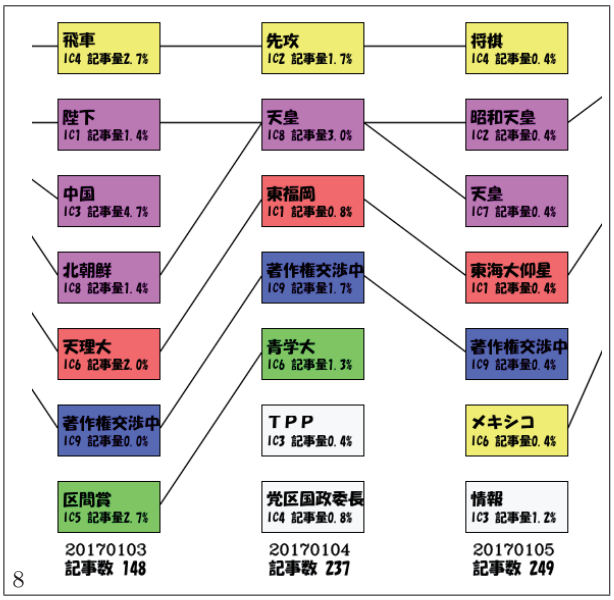

図 2: 話題変遷の可視化

5.4 時系列データの話題変遷の可視化

話題間の類似性を用いた時系列文書の話題の変遷を図 1 に示す. 図 1 より, 話題の分離, 話題の統合, 話題の継続及び新規話題が確認できる. 類似性がある話題間を同色に配色し, 類似性がない話題は白色で示している. 数日にわたり話題が継続している様子が確認できる. また各日ごとに新規の話題が出現している様子が確認できる.

類似性の有無を判定するための閾値の設定により, リンクの有無が大きく変化する. 閾値を高くすると, 話題の類似性が検出されず, 話題の分離, 話題の統合共に減少し, 新規話題が増える. 対して, 閾値を低くしすぎると, 話題間の類似性の質が低下する. 類似性判定の閾値の設定方法が課題であるといえる.

5.5 特定の単語による話題追跡

話題追跡の一例として, 特定の単語による話題追跡を紹介する. 特定の単語を「青学」とし, 青学に関連する話題を追跡する. 1 日ごとの記事で「青学」という単語出現数と, 独立話題分析を用いた話題追跡を比較する.

5.5.1 単語出現数

1 日ごとの記事データに特定の単語「青学」が含まれている日付と回数をグラフ化する. 単語出現率により青学が記事になった大まかな回数の把握が可能になる. 「青学」の単語出現数の推移を図 3 に示す. 単語出現数の推移の分析では, 特定の単語が含まれる記事がどのタイミングで出現したか, 分析することが可能である. しかし, 話題の内容については分析, 追跡することは容易でない.

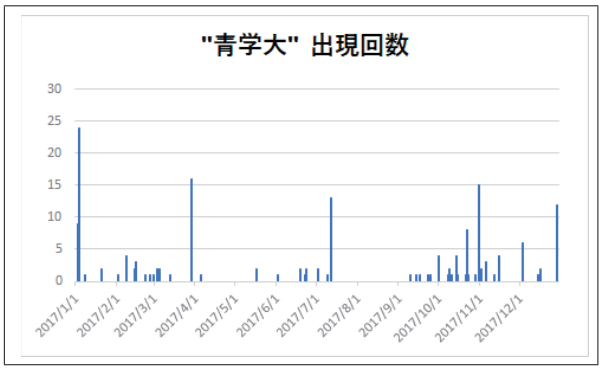

図 3: 2017 年「青学」単語出現数

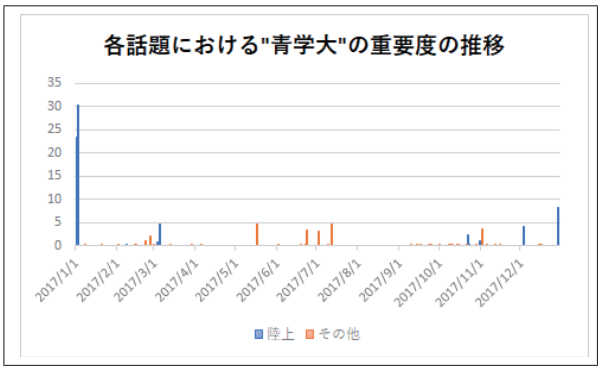

図 4: 独立話題分析を用いた話題抽出結果

表 2: “青学” を含む記事一覧の一例

月	日	見出し
1	3	ニューイヤー駅伝：旭化成、１８年ぶりＶ
1	3	箱根駅伝：青学大、往路３連覇
1	3	全国高校ラグビー：御所実－石見智翠館
1	4	箱根駅伝：３連覇３冠、青学時代
1	6	★ハンドボール：世界選手権　５戦全敗
1	7	ゴルフ：ＳＭＢＣシンガポール・オープン
…		
3	30	プロ野球：２０１７年選手名鑑セ・リーグ　広島
3	30	プロ野球：２０１７年選手名鑑セ・リーグ　巨人

5.5.2　特定の単語が含まれる記事の抽出

データセットから特定の単語が含まれている記事を抽出する．特定の単語「青学」を含む記事一覧の一例を表 2 に示す．

5.5.3　独立話題分析を用いた話題追跡

独立話題分析を用いた話題追跡結果をグラフ化する独立話題分析の各話題における単語の重要度の値の推移により，話題の追跡を行う．各話題における「青学」の重要度の推移を図 4 に示す．

グラフから単語出現数の推移と同様に，話題がどのようなタイミングで出現したか，分析することが可能である．また，独立話題分析による話題間の類似性を取り入れることで，特定の単語の検索でどのような話題が含まれているか分析することが可能である．グラフからは，陸上，その他の話題が含まれていて，それぞれの話題が出現したタイミングと共に可視化することができる．重要度の高い 1 月の話題は陸上の話題であり，表 1 から箱根駅伝の記事であることも確認できる．「青学 3 連覇」のように，話題を構成する上で影響を大きく与える単語の場合重要度の値も大きくなる．

5.5.4　比較結果

単語出現数のグラフでは，単語が出現した回数の推移を示している．しかし，単語出現数が多くても，話題に取り上げられないことも現実では多くある．したがって単語出現数から話題追跡を行うことは容易でない．独立話題分析を用いた話題抽出結果では，話題の類似性，特に話題の復活を捉えることにより，箱根駅伝，全日本駅伝等の駅伝関連の話題の発生を追跡できる．また，独立話題分析による話題追跡結果は単語の出現数に依存しない．「青学」の単語による話題追跡結果では,3 月 30 日の単語出現数が年間で 2 番目の 17 回である．しかし，独立話題分析による単語の重要度は,0 に近い．表 1 よりこの記事は，毎日新聞のコラムの 1 つである，プロ野球選手名鑑という記事である．記事は，日本野球球界 12 球団分あり，青学出身の選手の紹介で「青学」という単語が使用されていることが記事の見出し及び記事本文の閲覧により確認できる．新聞のコラムで，世間的な話題となっていないため，独立話題分析により抽出された重要度は低くなる．以上の結果から，独立話題分析を用いた話題追跡結果では，特定の単語から関連する話題の内容及び発生時期の追跡が可能である．また，単語の出現数に依存せず，記事の内容により，話題追跡ができる．

また，「青学」は単語の出現数が少ないため，話題の分離，統合は見られないが，単語出現数が多い単語には，同期間内で複数の話題に重要度を持つことも考えられる．

6.　まとめ

本稿では，独立話題分析を用いた時系列文書の話題変遷の可視化，及び話題追跡手法について述べた．本手法は独立話題分析に基づいており，話題分析の出力を利用することで可視化した．話題を期間ごとに配置し，類似度に基づいたリンクを張ることで，時系列文書における話題変遷を可視化した．また，話題追跡としては，特定の単語に関連する話題の追跡事例を紹介し，特定の単語に関連する話題の内容及び発生時期の追跡が可能である結果を示した．

今後の課題としては，閾値の設定方法がある．閾値が高いと話題の統合や分離，話題追跡結果がうまく検出されず，新規話題が増えてしまうことから，閾値の決定方法が課題である．またユーザによる評価実験を行う必要がある．

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)

[毎日新聞]　提供権者:毎日新聞社, 発売:日外アソシエーツ,CD-毎日新聞データ集

[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)