人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第11回)
SIG-AM-11-08

文書ストリームにおけるトピックダイナミクスの階層化ビジュアリゼーション

Hierarchical Visualization System of Topic Dynamics in Document Stream

澤井裕介 [1*]　熊野雅仁 [2]　木村昌弘 [2]
Yusuke Sawai[1]　Masahito Kumano[2]　Madahiro Kimura[2]

[1] 龍谷大学大学院理工学研究科電子情報学専攻
[1] Division of Electronics and Informatics, Ryukoku University
[2] 龍谷大学理工学部電子情報学科
[2] Department of Electronics and Informatics, Ryukoku University

Abstract: ソーシャルメディアの発達により WEB 上に大規模文書ストリームが多数出現しており，それらをわかりやすく整理，説明することが強く求められている．近年，文書ストリームにおけるトピックの融合・分離に着目したトピックダイナミクスが注目されている．しかし，従来は，トピックの活性度が考慮されていなかった．本研究では，新聞データを用いて日々のトピック間の関係や活性度の変化を視覚的に分析できる TimeLine を用いた階層的可視化システムを提案する．

1 はじめに

日々刻々と変化する世界の情勢を把握することは，社会を生きる人々の重要な関心の対象といえる．これまで，世界の情勢は，新聞，ラジオ，テレビなどの主要メディアに携わる人々によって，限られた紙面，時間制限の中で情報の取捨選択せざるを得ない状況があった．しかし，近年，WEB 世界の発展により，世界情勢を伝える発信源は無尽蔵に増え続け，主要メディアで取り上げられなかった情報はもとより，ソーシャルメディアの発達により，これまで主要メディアの情報を知る側であった人々が意見を述べ，情報の発信源としても成長していることから，WEB 空間に文書ストリームが無数に出現している．また，ソーシャルメディアが報じる無数の情報は，主要メディアの報じる内容にまで影響を及ぼし始めているため，主要メディアやソーシャルメディアで生み出される無数のトピック間相互に，複雑に影響し合う関係が成立しており，ダイナミクスが存在し得ると予想される．

そのようなトピックのダイナミクスを捉えるためには，観測できる対象として，複数の文書ストリーム間の関係を捉え，それらをわかりやすく整理，分析できる環境を構築することが望ましいと思われる．ただし，文書ストリームにおけるトピックは一定ではなく，日々，変容しており，ひとつのトピックが分離して発展・活性化したり，複数のトピックが融合して活性化するなど，様々な変化・変動が起きている可能性がある．近年，トピックの動的な変化を扱う研究 [1] や，時間軸上で前後に位置するトピック同士の依存関係響に着目して時間展開を捉える研究 [2]，トピックの発生と消滅を捉える研究 [3] や，さらにはトピックの分離・融合を扱う研究 [4] など，トピックの時間依存性やトピック相互の関係に着目した研究が注目されている．

ただし，生活や文化，政治や経済などの一般的なトピックには，政治の場合，例えば税金，外交，防衛など，大局的に安定して存在し続けるトピックが存在する．これらのトピックは，完全に消滅するとは考えにくいと思われる．一方，沖縄問題，尖閣諸島，オスプレイなどは，異なるトピックと言えるが，日々，活発に話題になったり，沈静化したりと，変動する傾向があるだけでなく，安定して存在する防衛トピックや，外交トピックのいずれとも関係がある．ここで，一年間で安定して存在する話題を年間主要トピック，日々，変動の大きい話題をデイリートピックとしたとき，トピックを階層的に捉えつつ，時間変化を捉える方法が考えられる．その観点において，デイリートピックは，無から発生したり，完全に消滅するのではなく，安定した複数の年間主要トピックと影響し合いながら，活性度が高まったり，沈静化しているにすぎないと考えることができる．また，デイリートピック同士も，相互に影響を及ぼし合いながら，分離して発展したり，相

*連絡先： 滋賀県大津市 瀬田大江町横谷 1-5
龍谷大学大学院理工学研究科
E-mail:t14m009@mail.ryukoku.ac.jp

互の依存関係に応じて融合するようなトピックであると考える．我々は，複数存在する文書ストリームに対して，安定した年間主要トピックと，変化・変動が起きやすいデイリートピックの関係を，包括的に捉えることでトピックダイナミクスを分析できる可能性に期待している．

そこで本研究では，生活や文化，政治や経済などの主要なトピックが含まれる新聞データを用いて，トピックを階層的に捉え，年間主要トピックとデイリートピックとの関係を可視化しつつ，デイリートピック間の関係や活性度の変化を時間軸に沿って視覚的に分析できる TimeLine を用いた階層的可視化ビジュアリゼーション法を提案する．本稿では，本研究の第一歩として，主要メディアのトピックを階層的に Timeline 上で捉えた際の視覚的分析に関する可能性を探るため，毎日新聞データセットを用いた実データによる実験で，提案システムの有効性を示す．

ただし，この方法では，トピックが分離や融合して話題が活性化したのか，それとも沈静化したのかなどを分析することができない．しかし，もし分離・融合とともに活性度がわかれば，より詳細に変動の様子を分析できる可能性がある．また，年間を通して安定して存在する年間主要トピックでも，活性度は変化している可能性があり，デイリートピックとの関係度の強さも変動している可能性がある．このため，年間主要トピックがどの日に活性化しているかや，年間主要トピックとデイリートピックの関係を可視化できれば，さらに詳細に変動分析が可能になることが期待される．本研究では，ダイナミクスの一面として，トピックの分離・融合だけでなく，活性度を可視化し，さらに年間主要トピックとデイリートピックとの関係度を可視化することで，文書ストリームのダイナミクスにおいて，複数の面から変動を分析することができるトピックダイナミクスの階層化ビジュアリゼーション法を提案する．

2 提案法

本稿では、文書群の時系列データ（文書ストリーム）として 1 年間の新聞記事群を考え、そのトピックダイナミクスの可視化法を提案する．

2.1 入力データ

ある年の新聞記事の全体（文書ストリーム）

$$\mathcal{D} = \bigcup_{t=1}^{T} D_t$$

を入力データとする．ここに、T は文書ストリームの総日数 (365 または 366) であり、D_t は第 t 日における記事全体の集合

$$D_t = \{d_{t,n} \,|\, n = 1, \cdots, N_t\} \ (t = 1, \cdots, T)$$

である．ただし、$d_{t,n}$ は第 t 日の第 n 記事であり、N_t は第 t 日における記事の総数である．各記事 $d_{t,n}$ は、形態素解析を行い、単語頻度ベクトル

$$\boldsymbol{x}_{t,n} = (x_{t,n,1}, \cdots, x_{t,n,V})$$

により BoW(bag-of-words) 表現する．ここに、各 $x_{t,n,i}$ は、想定する語彙集合 $\{voc_1, \cdots, voc_V\}$ に対し、文書 $d_{t,n}$ における語彙 voc_i の出現回数である．V は、想定する語彙の総数である．

2.2 年間主要トピックスの抽出

文書ストリーム $\mathcal{D}$ における年間レベルでの主要トピックの出現と消滅のダイナミクスを調べるために、文書群データ $\{D_t \,|\, t = 1, \cdots, T\}$ を多重トピックを考慮した文書の確率的生成モデルである HDP-LDA(hierachical Dirchlet Process - Latent Dirichlet Allocatio)[5][1] によりモデル化する．

各 t に対して、第 t 日の記事群 D_t を、それに属するすべての記事を単純につなぎ合わせて一つの長い文書と考え、単語頻度ベクトル

$$\boldsymbol{X}_t = (X_{t,1}, \cdots, X_{t,V})$$

により BOW 表現する．ここに、各 $X_{t,i}$ は

$$X_{t,i} = \sum_{n=1}^{N_t} x_{t,n,i}$$

である．そして HDP-LDA モデルに基づいて、観測データ $\{\boldsymbol{X}_t \,|\, t = 1, \cdots, T\}$ に対する、潜在トピック集合

$$Y = \{y_1, \cdots, y_L\}$$

および、各潜在トピック $y \in Y$ の下での単語生成ベクトル

$$\boldsymbol{\theta}_y = (\theta_{y,1}, \cdots, \theta_{y,V})$$

を、それぞれ推定する．我々は、各 $y \in Y$ を文書ストリーム $\mathcal{D}$ の**年間主要トピック**と呼び、それら年間主要トピックスを抽出し分析する．

まず、文書ストリーム $\mathcal{D}$ の年間主要トピック $y \in Y$ が、第 t 日にどのくらい活発であったかを、事後確率を用いて、

$$f_y(t) = P(y \,|\, \boldsymbol{X}_t)$$

により測定する．我々は、$f_y(t)$ を年間主要トピック y の第 t 日における**活性度**と呼ぶ．t に関する $f_y(t)$ の変

動を調べ、年間主要トピック y のダイナミクスを分析する.

また、各 $y \in Y$ に対し、潜在トピック y の下での単語生成ベクトル $\boldsymbol{\theta}_y$ においてランキングを行うことによって y と関係がより深い単語を抽出することにより、年間主要トピック y を説明する.

2.3 デイリートピックスの抽出

文書ストリーム $\mathcal{D}$ における日レベルでのトピックについて調べるために、各 t に対して、第 t 日の文書群 $D_t = \{d_{t,k} \,|\, k = 1, \cdots, K_t\}$ を多重トピックを考慮した文書の確率的生成モデルである HDP-LDA によりモデル化する. そして、HDP-LDA モデルに基づいて、観測データ $\{\boldsymbol{x}_{t,k} \,|\, k = 1, \cdots, K_t\}$ に対する、潜在トピック集合

$$Z_t = \{z_{t,1}, \cdots, z_{t,K_t}\}$$

および、各潜在トピック $z \in Z_t$ の下での単語生成ベクトル

$$\boldsymbol{\phi}_{t,z} = (\phi_{t,z,1}, \cdots, \phi_{t,z,V})$$

を、それぞれ推定する. 我々は、各 $z \in Z_t$ を文書ストリーム $\mathcal{D}$ における第 t 日の**デイリートピック**と呼び、それらデイリートピックスを抽出し分析する.

まず、任意の t に対して、第 t 日の各デイリートピック $z \in Z_t$ を次の 2 つのやり方で説明する.

1. 事後確率 $P(z \,|\, \boldsymbol{x}_{t,k})$ が高い記事 $d_{t,k}$ を抽出する.
2. 単語生成ベクトル $\boldsymbol{\phi}_{t,z} = (\phi_{t,z,1}, \cdots, \phi_{t,z,V})$ において要素の値が大きい語彙 voc_i を抽出する.

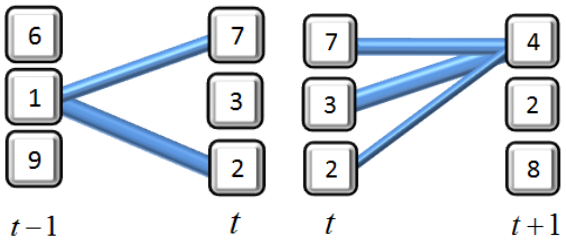
また、デイリートピック $z \in Z_t$ がどのくらい活発であったかを、事後確率を用いて、

$$f_z(t) = \sum_{n=1}^{N_t} P(z \,|\, \boldsymbol{x}_{t,n})$$

により測定する. 我々は、$f_z(t)$ をデイリートピック $z \in Z_t$ の**活性度**と呼び、デイリートピックスの活性度を分析する.

次に、第 t 日のデイリートピック $z \in Z_t$ がどの年間主要トピック $y \in Y$ と関係しているかを、単語生成ベクトル $\boldsymbol{\phi}_{t,z}$ と $\boldsymbol{\theta}_y$ のコサイン類似度で測定する. 我々は、その値をデイリートピック $z \in Z_t$ と年間主要トピック $y \in Y$ の**関係度**と呼び、デイリートピックスと年間主要トピックスの関係度を分析する.

次に、第 t 日のデイリートピック $z \in Z_t$ が、翌日である第 $t+1$ 日のデイリートピック $z' \in Z_{t+1}$ とどのように関係しているかを、単語生成ベクトル $\boldsymbol{\phi}_{t,z}$ と $\boldsymbol{\phi}_{t+1,z'}$ のコサイン類似度で測定する. 我々は、その値をデイリートピック $z \in Z_t$ とデイリートピック $z' \in Z_{t+1}$ の**関係度**と呼び、隣接する日々のデイリートピックス間の関係度を分析する.



(a) 分離　　(b) 融合

図 1: デイリートピックの分離と融合

2.4 トピックのダイナミクス

1 年間の文書ストリーム全体に対し，HDP-LDA[5][1] を適用すると，年間を通じた潜在トピックが得られる. ただし，この年間を通じた潜在トピック（年間主要トピック）は，年間を通じて安定して存在するトピックの一面を捉えていると考えられるものの，日々の変動を捉えてはおらず，この年間主要トピックだけに着目してもダイナミクスを捉えることはできない.

一方，一日ごとに多数の情報源（新聞であれば多数の記事）に対して HDP-LDA を適用すると，一日ごとの潜在トピック（デイリートピック）が得られる. デイリートピックは，日々変動する話題内容の変容を捉えている可能性があるだけでなく，日ごとに変化し得る潜在トピック数の変動も捉える可能性がある. 図 1 は，ある日に得られたデイリートピックを番号で表し，次の日に得られたデイリートピックとの関係を線で結んだ様子を示しているが，線の太さに関係の強さを割り当てている例である. このような潜在トピックの可視化を実現すれば，図 1(a) のように，第 $t-1$ 日のデイリートピック $z_{t-1,1}$ が，第 t 日のデイリートピック $z_{t,7}$ と $z_{t,2}$ に分かれるようなトピックの分離を捉え得る可能性がある. また，図 1(a) のように，第 t 日のデイリートピック $z_{t,7}$, $z_{t,3}$, $z_{t,2}$ が，第 $t+1$ 日のデイリートピック $z_{t+1,4}$ に集中するようなトピックの融合を捉える可能性もあるため，デイリートピック間のダイナミクスを捉える可能性がある.

3 提案システムデザイン

提案システムの概観を図 2 に示す. 図 2 のように，提案システムは，四つの View で構成される. 基本となる View A は，年間主要トピック $y \in Y$ の全体をタイムライン上に可視化するものである. View B は，View

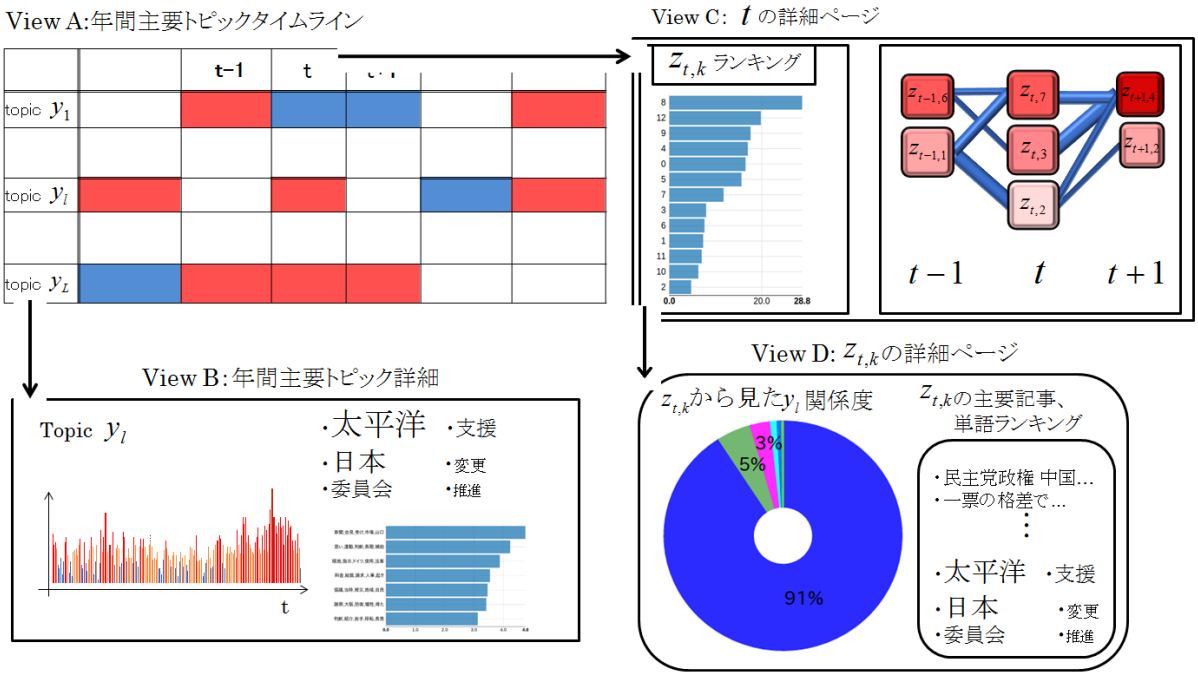


図 2: 四つの View で構成される提案システムの概観

A に表示された一つの年間主要トピック $y \in Y$ の詳細を可視化するものである．View C は，View A に表示されたある日 t に対応する活性度の高いデイリートピックと，日 $t-1$ と日 $t+1$ に対応する活性度の高いデイリートピックとの関係度の強さを可視化するものである．View D は，個々のデイリートピックで出現確率の高い単語のランキング上位や，デイリートピック $z \in Z_t$ と年間主要トピック $y \in Y$ の関係度の強さを可視化するものである．次に，これら四つの View について，個々に詳細を説明する．

3.1 年間主要トピックタイムライン:View A

提案法では，年間主要トピック $y \in Y$ に関して，日単位で活性度 $f_y(t)$ を算出することができる．活性度は，高い状態を赤，中間を白，低い状態を青で表す．図 2 の View A は，その可視化の様子を示したものである．この View A により，ユーザは，どの年間主要トピックがいつ活性化しているかを確認することができる．

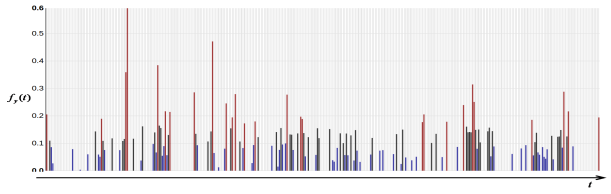


図 3: 年間主要トピック $y \in Y$ における各日の活性度

また，View A 上の年間主要トピック名を押すと，View B により，年間主要トピックの詳細を確認することができ，View A 上の日名を押すと，日 t に含まれるデイリートピックに関する詳細を知ることができる．

3.2 年間主要トピックの詳細：View B

一つの年間主要トピックに焦点を当て，より詳しい情報を提示する View である．より詳しい情報として，横軸を日とした一年，縦軸を活性度とした図 3 のグラフを通じて，活性度のより細かい変化を確認することができる．

また，図 2 の View B において，右上部にフォントサイズの異なる単語を確認することができる．これは，年間主要トピック $y \in Y$ において，Bag of words 表現された単語の出現確率の高さをフォントサイズの大きさで表現したものである．これにより，年間主要トピック $y \in Y$ と関連の高い単語のランキング上位を確認することができる．

また，図 2 の View B において，右下部にある棒グラフを拡大したものが図 4 である．これは，年間主要トピック $y \in Y$ とある日 t のデイリートピック $z \in Z_t$ との関係度ランキングを表しており，ランキングの上位が降順に整列されている．これにより，View B を用いて注目している年間主要トピック $y \in Y$ が，ある日 t のデイリートピック $z \in Z_t$ とどのような関係にあるかを確認することができる．また，図 4 のように，各

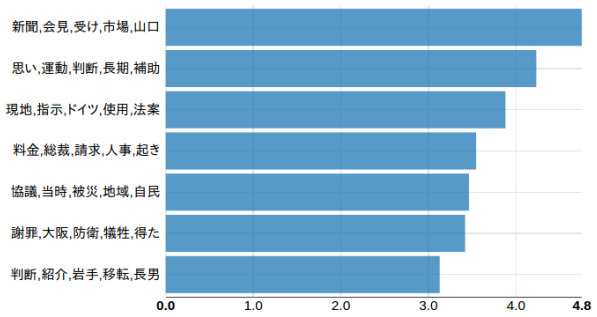


図 4: 年間主要トピック $y \in Y$ からみたデイリートピックス Z_t の関係度ランキング

デイリートピック $z \in Z_t$ の棒グラフの左側には単語生成ベクトル $\boldsymbol{\phi}_{t,z} = (\phi_{t,z,1}, \cdots, \phi_{t,z,V})$ において要素の値が大きい語彙 voc_i の Top 5 が表示されている．これにより，活性度の高い $z \in Z_t$ が，どのような単語に強い関係を示すかがわかり，その $z \in Z_t$ が表す潜在的トピックの内容を調べる足がかりとなる．

3.3 第 t 日の詳細：View C

図 2 の View C では，View A に表示された第 t 日のデイリートピックス Z_t に関する詳細が確認できる．図 2 の View C の左側にある棒グラフを拡大した一例が図 5 である．これは，第 t 日に抽出されたデイリートピック $z \in Z_t$ を活性度 $f_z(t)$ によってランキングを行い，降順に可視化したものである．これにより，各デイリートピック $z \in Z_t$ の活性度の値がわかり，活性度のランキングもわかるため，どのデイリートピックがどの程度活性化しているかを確認することができる．

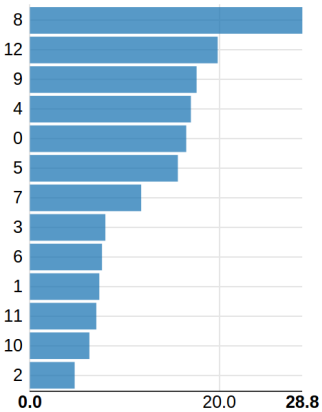


図 5: デイリートピック $z \in Z_t$ の活性度ランキング

また，図 2 の View C の右側では，第 t 日だけでなく，第 $t-1$ から第 $t+1$ 日までのデイリートピック間の関係度の変化がわかるだけでなく，日ごとにデイリートピック Z_t の活性度に関するランキング情報が可視化されるため，デイリートピック間の関係度と活性度の変化を同時に分析することができる．この可視化については，3.5 で説明する階層的トピックダイナミクス分析と関連が深いため，3.5 で詳細に説明を行う．

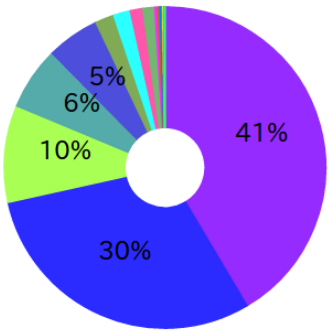


図 6: デイリートピック $z \in Z_t$ と年間主要トピックス Y との関係度

3.4 各デイリートピックの詳細：View D

図 2 の View D は，View C において，一つのデイリートピックの詳細情報を見るため，注目するデイリートピック $z \in Z_t$ を押した際に表示されるものである．図 2 の View D の右側には，注目するデイリートピック $z \in Z_t$ での voc_i を確認することができる．これにより，デイリートピックの内容を調べる足がかりとなる．また，図 2 の View D の左側にある円グラフを拡大したものが図 6 である．図 6 は，各デイリートピック $z \in Z_t$ と各年間主要トピック $y \in Y$ の関係度を円グラフで表現したものである．これにより，デイリートピックがどの年間主要トピックと関係が強いかを知ることができる．また，異なるデイリートピックの円グラフを図 2 の View C の右側の部分で同時に表示することができる．この機能を用いれば，図 7 のように，表示することで，数日間のデイリートピック間の関係度の変化や活性度の変化を同時に分析できるだけでなく，さらに，個々のデイリートピック $z \in Z_t$ と年間主要トピック $y \in Y$ との関係度を円グラフを通じて確認することができるため，デイリートピックの分離・融合に関するダイナミクスだけでなく，階層的に活性度や関係度のダイナミクスを分析できる可能性がある．

3.5 階層的トピックダイナミクス分析

図 7 は，図 2 の View C 右側にある数日間のデイリートピックタイムラインにおいて，いくつかのデイリートピックに関する年間主要トピックとの関係率を示す円グラフを同時に可視化した例である．

この可視化により確認できることは，まず，第 $t-1$ 日から第 $t+1$ 日までの各日に抽出されたデイリートピック $z \in Z_t$ の個数が変化する様子である．図 7 にお

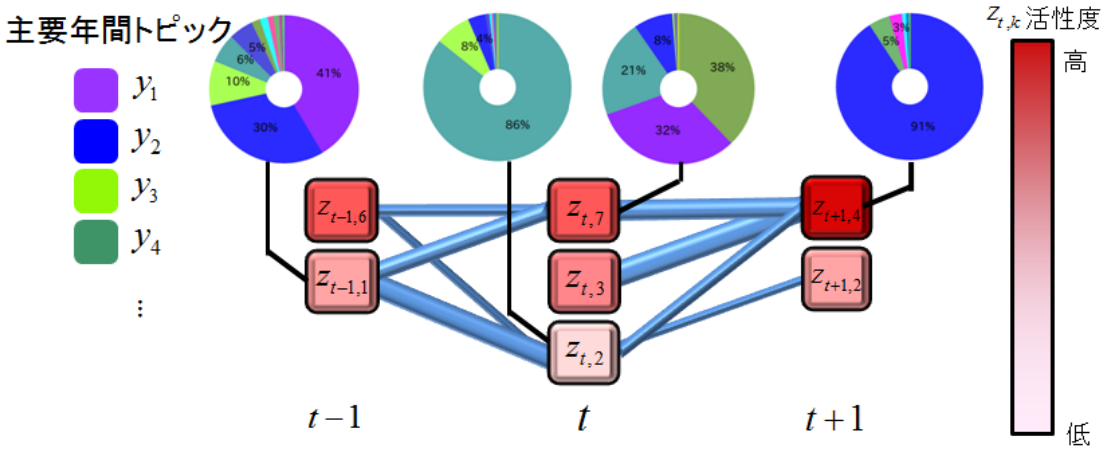


図 7: デイリートピックのタイムラインを用いた階層的トピックダイナミクス分析

いて，第 $t-1$ 日には二つのデイリートピック z があり，第 t 日には三つのデイリートピック，第 $t+1$ 日には二つのデイリートピックがあることが見てとれる．また，各デイリートピックは，日ごとに活性度ランキングに基づいて縦方向に降順で整列されているため，各日の最も上に位置する $z \in Z_t$ が，その日に活性度が最も高いデイリートピックとなる．さらに，各 $f_z(t)$ は，図 7 右にあるような白から赤の色で着色されているが，これは，活性度の値を示している．これにより，たとえ，ある日の活性度が最も高いデイリートピックであっても，他の日と比較すると，活性度に差があることを視認できる．例えば，第 $t-1$ 日では，$z_{t-1,6}$ が最も活性度が高く，第 t 日では，$z_{t,7}$ 第 $t+1$ 日では，$z_{t+1,4}$ が最も活性度が高いデイリートピックであることが見てとれるが，第 $t-1$ 日の $z_{t-1,6}$ や第 t_2 日の $z_{t,7}$ よりも，第 $t+1$ 日では，$z_{t+1,4}$ の赤色の彩度が最も高いことが見てとれる．つまり，活性度の変化をより詳しく視認することができる．

次に，第 $t-1$ 日のデイリートピック $z_{t-1,1}$ は，第 t 日で活性度が最も高い $z_{t,7}$ と，活性度が低い $z_{t,2}$ に分離している可能性があることや，第 t 日の特に $z_{t,7}$ と $z_{t,3}$ が，第 $t+1$ 日では $z_{t+1,4}$ に融合していることがわかる．$z_{t+1,4}$ は，第 $t-1$ 日から第 $t+1$ 日の中で，最も彩度の高い赤を示していることから，何かが起きている可能性を期待させる．このように，活性度を用いると，分離や融合の観点だけでなく，より詳しくデイリートピックの変化を捉えることができる可能性がある．

さらに，図 7 では，デイリートピックの分離・融合に関係していると見なした第 $t-1$ 日の $z_{t-1,1}$，第 t 日の $z_{t,7}$ と $z_{t,2}$，第 $t+1$ 日の $z_{t+1,4}$ に関するデイリートピックと年間主要トピックとの関係率を示す円グラフを選択的に同時表示した様子を示している．円グラフの色は，図 7 左端にあるように，年間主要トピック

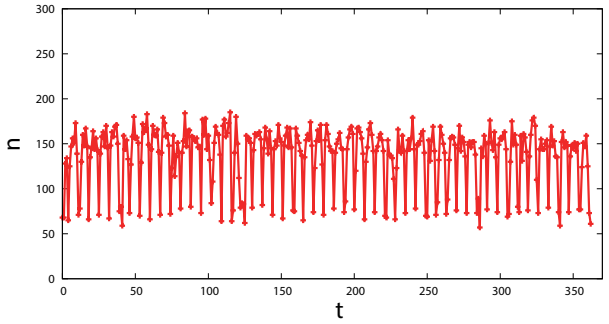


図 8: 各 t の記事数 n

$y \in Y$ を識別するために割り当てた色であり，図 7 の各円グラフの変遷から，分離・融合に携わるデイリートピックと関係度の高い年間主要トピックが変化している様子も視認できることがわかる．このように、提案可視化法では，年間主要トピックとデイリートピックを階層的に捉えながら活性度や関係度の動的な変化を視覚的に分析できることがわかる．

4 実施例

4.1 実験データ

本研究では実験データとして，毎日新聞データベースより，2013 年 1 月 1 日から 2013 年 12 月 31 日の新聞記事の中で 1 面，2 面，3 面，経済面，社会面の記事を文書ストリームとして使用した．1 日あたりの記事数 n のグラフを図 8 に示す．これによりおおよそすべての日において偏りなく記事が書かれていることが分かる．また，これらの記事において，1 日に 1 回以上出現している単語という条件のもと BOW 表現に変換した．その語彙の総数は，1,333 となった．

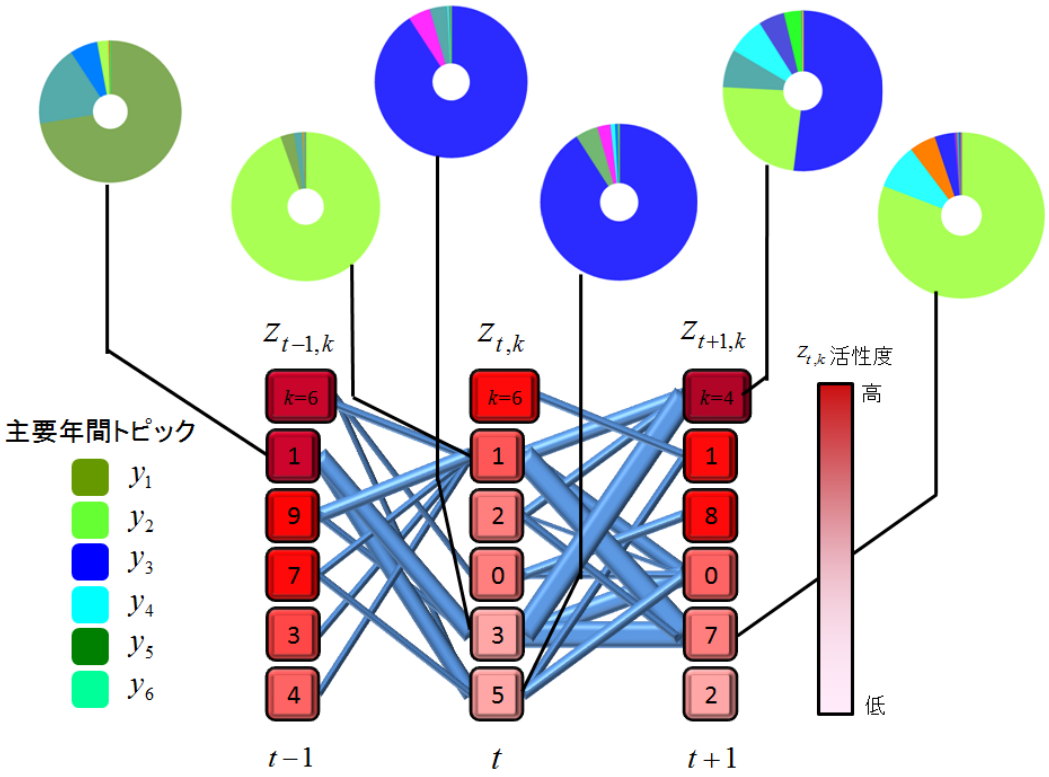


図 10: t=7/21 のデイリートピックタイムライン

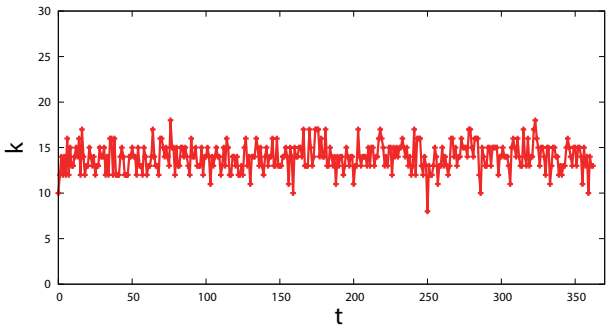


図 9: デイリートピック数 k の変動

4.2 実験

2013 年の実験データに対し，潜在トピック推定法を適用したところ，抽出された年間主要トピック $y \in Y$ の総数は 18 個となった．また，各 t 日において，デイリートピックを推定したところ，各 t 日のディリートピック数 K_t は，図 9 に示すような結果となった．図 9 の横軸は t となっており，縦軸は各 t 日のディリートピック数となっている．すべての日において，デイリートピックが一定数以上生成されていることや，日によって数が変動していることが確認できる．

2013 年のデータを用いて可視化システムの View A を表示させたのち，ディリートピックの詳細として，t=7 月 21 日を選択し，View C を表示した．このとき，階層的トピックダイナミクス分析が可能なディリートピックタイムラインを表示させたものが図 10 である．2013 年，7 月 21 日は，参院選挙の開票日であり，第 $t-1$ 日はその前日，第 $t+1$ 日は化参院選挙の開票日直後の日となる．図 10 より，第 t 日のディリートピック $z_{t,1}$ が第 $t+1$ 日のディリートピック $z_{t+1,4}$ と $z_{t+1,7}$ に分離している様子が見てとれる．ただし，ディリートピック間の関係度の強さが可視化されているだけでなく，活性度ランキングと活性度の値が可視化されているため，第 t 日のディリートピック $z_{t,1}$ は，活性度ランキング上トップで，活性度もかなり高い第 $t+1$ 日のディリートピック $z_{t+1,4}$ と結ばれている．しかし，第 t 日のディリートピック $z_{t,1}$ は，むしろ活性度ランキングが低めの第 $t+1$ 日のディリートピック $z_{t+1,7}$ と関連が強いことがわかる．ただし，第 t 日のディリートピック $z_{t,1}$ と第 $t+1$ 日のディリートピック $z_{t+1,7}$ は，活性度を示す色がほぼ同色であることから，活性度は変化していないが，$z_{t,1}$ や $z_{t+1,7}$ よりも，より活性度の高いディリートピック（例えば $z_{t+1,4}$）が第 $t+1$ 日に現れ，上位を占めてたと解釈できる可能性がある．

ところで，第 $t+1$ 日のディリートピック $z_{t+1,4}$ は，第 t 日のディリートピック $z_{t,1}$ と $z_{t,3}$ が融合したものであると解釈できる可能性がある．また，これらの $z_{t,1}$ と $z_{t,3}$ に関して，年間主要トピックとの関係率を示す

円グラフを表示させたところ，図 10 より，$z_{t,1}$ は，ほぼ黄緑の年間主要トピックと関係し，$z_{t,3}$ は，ほぼ青の年間主要トピックと関係していることがわかる．さらに，$z_{t,1}$ と $z_{t,3}$ が融合したと解釈した $z_{t+1,4}$ の円グラフを確認すると，その主要な年間主要トピックは，ほぼ黄緑の年間主要トピックと青の年間主要トピックであることが見てとれるため，$z_{t,1}$ と $z_{t,3}$ の融合したものが $z_{t+1,4}$ であるという解釈をより裏付けている可能性が期待される．ディリートピック $z_{t,1}$，$z_{t,3}$，$z_{t+1,4}$ のそれぞれと関連の深い第 $t+1$ 日の新聞記事をランキングしたことろ，以下のような記事と関連が高かった．$z_{t,1}$ は，エジプトでの武装勢力の攻撃に関する記事や，中国の影響力を懸念する記事，中国でのテロ行為．$z_{t,3}$ は，投開票日の話題や，与党と野党の攻防．$z_{t+1,4}$ は，自民圧勝，ねじれ解消，アベノミクスが指示されたという解釈が報じられた記事であった．この事例の解釈は，ユーザに委ねるものの，提案可視化法では，ディリートピックの分離・融合の観点だけでは捉えきれない，より詳細なトピックダイナミクスの分析が可能になる点で，提案法の有効性が示唆されていると思われる．

5 まとめ

本研究では，文書ストリームに対し，タイムライン上で，日々のデイリートピックの関係度を可視化することで，分離・融合するデイリートピック間のダイナミクスを分析するだけでなく，年間主要トピックとデイリートピックの階層的な関係を示しながら，トピックの活性度を可視化することで，より多くの観点からダイナミクスを視覚的に捉える文書ストリームのトピックダイナミクスにおける階層的ビジュアリゼーション法を提案した．提案法では，まず，文書ストリームに含まれる文書情報を 1 年単位で BOW 表現し，未知数の潜在的な多重トピックをパラメータ学習によって決定できる HDP-LDA を用いて推定する．ただし，1 年間すべての文書データに HDP-LDA を適用して年間主要トピックを推定し，日々の文書データに HDP-LDA を適用してデイリートピックを推定する．次に，年間主要トピック y に関して，第 t 日における活性度 $f_y(t)$ およびデイリートピック z に関して，活性度 $f_z(t)$ を求める．そして，年間主要トピックとデイリートピックの関係度を求める．このようにして得られた年間主要トピックとデイリートピックに関して階層的に文書ストリームにおけるトピックダイナミクスを活性度，関係度の観点から可視化し，視覚的分析が可能な環境を構築した．

実データとして，毎日新聞データベースを文書ストリームと見なし，階層化ビジュアリゼーション法を含む提案法の有効性を評価した．2013 年のデータを用いることにより，提案法では，まず，活性度をデイリートピックの活性度ランキング用い，さらに色を用いて活性度の値を可視化することで，デイリートピックの分離・融合を捉えるだけでなく，分離したデイリートピックが活性化したのか，沈静化したのか，また，融合したデイリートピックが活性化したのか，沈静化したのかという情報に加え，活性度の値はどの程度変化したのかがわかり，より詳しくデイリートピックの活性度の変化を分析できる可能性を示した．また，注目したデイリートピックの年間主要トピックとの関係率を示す円グラフを選択的，かつ複数同時に可視化することにより，分離・融合しているデイリートピックと，年間主要トピックの関係が，変わらない場合や変化することを確認することができ，文書ストリームのトピックダイナミクスを階層的に分析できる可能性も示し，提案法の有効性を示した．

参考文献

[1] D.M. Blei and J.D. Lafferty, “Dynamic topic models,” Proceedings of the 23rd International Conference on Machine Learning, pp.113–120, ICML ’06, ACM, 2006.

[2] A. Ahmed and E.P. Xing, “Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering,” Proceedings of the SIAM International Conference on Data Mining, SDM 2008, April 24-26, 2008, Atlanta, Georgia, USA, pp.219–230, 2008.

[3] A. Ahmed and E.P. Xing, “Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream,” UAI 2010, Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, July 8-11, 2010, pp.20–29, 2010.

[4] 佐々木謙太朗 C 吉川大弘 C 古橋　武 C “複数のトピックの時間的依存関係を考慮した時系列混合モデル C” 人工知能学会論文誌 C vol.30Cno.2Cpp.466–472C2015D

[5] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, “Hierarchical dirichlet processes,” Journal of the American Statistical Association, vol.101, pp.1566–1581, Dec. 2006.