情報処理学会論文誌　データベース　Vol.11　No.2　16–29 (July 2018)

ライフイベントの経験に有用なトピックの抽出と評価

武田 直人[1,a)]　関 洋平[2,b)]　森下 民平[3,c)]　稲垣 陽一[3,d)]

受付日 2017年12月10日, 採録日 2018年4月2日

概要：「出産」や「就職」などのライフイベントを経験することで，ユーザの興味や行動は変化する．ライフイベントの経験後にユーザが持つ興味や行動に関連したトピックを抽出し，それらのトピックを利用してブログ記事を提示することで，ライフイベント経験に有用な情報を提供することが可能になる．本研究では，ライフイベントを経験したユーザの投稿したブログ記事から，有用な情報を提示するためのトピックを抽出する手法を提案する．まず，時系列トピックモデルを用いて，各ブログ記事に出現するトピック分布を推定する．次に，ブログ記事内において，他のトピックと頻繁に共起するトピックをフィルタリングする．さらに，各トピックについて，ライフイベントの前後における投稿人数の増加と，ブログ記事内において，トピックが他のトピックと共起したブログ記事の数を計算し，トピックを抽出する．最後に，抽出したトピックの出現確率が高いブログ記事を提示する．「出産」,「就職」,「結婚」,「大学入学」,「子供の小学校入学」の各ライフイベントについて，各ライフイベントを経験したことのある 6 名の実験参加者により，提案手法を評価し，提案手法とトピックのバーストを考慮した手法を比較した．その結果，提案手法の有効性を確認した．

キーワード：ライフイベント，DTMs（Dynamic Topic Models），ブログ

Topic Detection and Evaluation Useful for Life Event Experience

Naoto Takeda[1,a)]　Yohei Seki[2,b)]　Mimpei Morishita[3,c)]　Yoichi Inagaki[3,d)]

Received: December 10, 2017, Accepted: April 2, 2018

Abstract: Users' interests and actions change as they experience life events such as "childbirth" and "finding employment." Detecting the topics relevant to users' interests or actions after the life event and extracting the blog posts from those topics make it possible to provide the useful information for life event experience. We propose a method to detect the topics for providing the useful information from blogs posted by users who have experienced life events. First, we estimate the topic distribution in each blog post using dynamic topic models. Second, we filter the topics which co-occurred frequently with other topics in each blog post. Third, we detect the topics by computing the increase in the number of posted users before and after the life event for each topic, and the number of blog posts of which the topic co-occurred with other topics in each blog post. Finally, we present the blog posts with high topic probabilities of the detected topics. We evaluated the proposed method by six research participants who have experienced each life event: "childbirth," "finding employment," "marriage," "university admission," and "child's admission to elementary school" and compared the proposed method with topic burstiness method. As a result, we confirmed the effectiveness of the proposed method.

Keywords: life event, DTMs (Dynamic Topic Models), blog

1. はじめに

「出産」や「就職」などのユーザの環境や習慣が変化す

[1] 筑波大学大学院図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba, Ibaraki 305–8550, Japan
[2] 筑波大学図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba, Ibaraki 305–8550, Japan
[3] 株式会社きざしカンパニー
kizasi Company, Inc., Chuo, Tokyo 103–0015, Japan

[a)] s1621623@u.tsukuba.ac.jp
[b)] yohei@slis.tsukuba.ac.jp
[c)] mimpei@kizasi.jp
[d)] inagaki@kizasi.jp

る出来事（ライフイベント）を経験することで，ユーザの興味や行動は変化する．たとえば，はじめて「出産」イベントを経験したユーザは，それまでに経験したことのない「育児」という行動をするようになる．このようなユーザを支援するために，ライフイベントを機に体験する新たな興味や行動について分析する研究がこれまでに数多く報告されている [4], [6], [15]．また，「出産」イベントで新しく母親となったユーザは，育児に関して，自身と同じ境遇のユーザの支援を求めることが報告されており [1]，ブログやSNS 上で同じ境遇のユーザが投稿した記事を参照することが報告されている [6], [12], [13]．そこで，本研究では，ライフイベントを経験したユーザの投稿から，これからライフイベントを経験するユーザが持つようになる興味や行動に関する有用な情報を含む投稿を提示することを目的とする．**表 1** に，有用な情報を含む投稿の例を示す．本研究では，表 1 の下線部の情報のような，具体的な体験談に基づいた，これからライフイベントを経験するユーザが知っておいた方が良い情報を含む投稿を，「有用な情報を含む投稿」と定義する．

ライフイベントを経験したユーザの投稿は，kizasi.jp [*1]やアメーバブログ[*2]などのブログサービス，Facebook [*3]やmixi [*4]のような SNS，Twitter [*5]などのマイクロブログ，Yahoo!知恵袋[*6]などの Q&A サイト，ウィメンズパーク[*7]などのライフイベントに関連した掲示板サイトなどで参照することができる．このうち，マイクロブログにおける投稿は，短文であることが多く，有用な情報が含まれにくい．また，Q&A サイトや掲示板サイトでは，投稿したユーザのライフイベントを経験したタイミングの特定が困難であり，ユーザの変化した興味や行動を明らかにすることは難しい．なお，出産や育児に関する掲示板サイトのウィメンズパークでは，子供の月齢別の悩みを投稿できる掲示板が提供されているため，ライフイベント後のユーザの悩みとその回答を提示できる可能性がある．しかし，悩みに関する掲示板の投稿には，ネガティブな投稿や，個人依存の強い投稿が多く，子供の具体的な成長過程や，性格の変化を示すような有用な情報を含むポジティブな投稿を提示することが難しい．また，「就職」や「結婚」などの，他のライフイベントについては，ライフイベント後の悩みの変化を追跡できる掲示板は存在しなかった．これに対して，ブログサービスや SNS における投稿では，ユーザの興味や行動の変化を追跡でき，長文の投稿も多いことから，様々なライフイベントで変化した興味や行動に関する有用な情報を含む投稿を提示することが可能である．本研究では，データの収集がしやすく，かつ，匿名性が高いため，ライフイベントに関する率直な悩みと，ポジティブな体験談の両方が投稿されやすいと考え，ブログを利用し，有用な情報を含む投稿を提示する．

表 1　有用な情報を含む投稿の例
Table 1　Examples of posts including some useful information.

ライフイベント	投稿の一部
「出産」	... 今のところはまだ授乳中にガブリとやられた事はありません．時々じわ～っと噛みながら私の目を見て笑ってます．**そんなときは「嫌がっても鼻をつまむ！」と言う助産師さんのアドバイスを実行してます．2～3 度繰り返すと「噛むと嫌なことがある」と赤ちゃんが学習して噛まなくなるんだって．**どうやら効果があるのではないかと思ってます．...
「結婚」	... 都民共済ブライダルプラザに行ってきました．休日だったこともあり，ものすごく混んでました．各種受付も，番号札を取ってから 30 分ぐらい待つ感じでした．店内には，結婚式場のパンフレットコーナー，結婚指輪コーナー，新婚旅行受付コーナー，引き出物コーナーなどがあり，どれも共済会員割引で申し込みできます．**披露宴会場や生花でのブーケも普通と比べて格段に安いけど，何よりも激安なのがウエディングドレス．10 倍は違います（笑）**私達は，当初の見積もりよりどんどん高くなっててビビッてるので，共済ブライダルプラザの価格を見てため息が出てしまいました．初めからここを知ってたら・・・.. ..

一方で，既存のブログサービスで提供されている，クエリによるブログ記事の検索では，有用な情報を含む投稿を検索することは難しい．たとえば，「育児」に関する有用な情報を含む投稿を検索する場合，「育児」という単語で検索すると，育児の制度に関する投稿や，育児への意気込みが書かれた投稿が大量に検索され，有用な情報が含まれるブログ記事を検索することはできない．また，kizasi.jp では，検索クエリと共起しやすい単語によるクエリ拡張を提供しており，「育児」という単語では，「仕事」，「家事」，「子供」などの単語が拡張される．これらの単語を用いて検索した場合も，育児の制度に関する投稿や，育児への意気込みが書かれた投稿が検索され，有用な情報を含むブログ記事やユーザの体験談が書かれたブログ記事を検索することはできなかった．

このような背景をふまえ，本研究では，ライフイベントを経験したユーザのブログ記事に出現するトピックを利用し，有用な情報を含むブログ記事を提示する．これにより，たとえば，「育児」に関する有用な情報を含むブログ記事を

[*1] http://kizasi.jp/
[*2] https://ameblo.jp/
[*3] https://www.facebook.com/
[*4] https://mixi.jp/
[*5] https://twitter.com/
[*6] https://chiebukuro.yahoo.co.jp/
[*7] http://women.benesse.ne.jp/

提示する場合，「育児」という単語ではなく，育児と関連が強く，ユーザの具体的な体験談に現れやすい「離乳食」や「授乳」などの単語集合を高い割合で含むブログ記事を，まとめて提示することができる．しかし，トピックモデルで推定するトピック集合には，これからライフイベントを経験するユーザにとって，有用でないトピックが大量に含まれる．提案手法では，有用な情報を含むブログ記事に出現するトピックは，ライフイベントを機に変化したユーザの興味や行動の変化を反映しており，かつ，ライフイベントと関連の深いトピックと仮定し，これらの特徴を持つトピックを抽出する．

具体的には，まず，ライフイベントを経験したユーザの興味や行動は，投稿するブログ記事のトピックとして出現すると考え，ライフイベントを経験したユーザの投稿したブログ記事を，各ユーザがライフイベントを経験した月からの経過月ごとに分割し，時系列トピックモデルの DTM（Dynamic Topic Models）[2] を適用する．これにより，各経過月のブログ記事に出現するトピック分布を推定する．次に，ブログ記事内において，ほとんどのトピックと共起するトピックはノイズとしてフィルタリングする．続いて，ライフイベントを経験したユーザの興味や行動の変化を反映したトピックを抽出するために，ブログ記事に対するトピックの出現確率を用いて，そのトピックに対する各経過月の投稿人数[*8]を推定し，ライフイベント前後における投稿人数の増加およびライフイベント後の特定の時期における投稿人数の増加を考慮したスコアリングを行う．さらに，多くのユーザが，共通する複数のトピックを同一の時期に共起させて書いた記事は，ライフイベントと関わりが深く，これからライフイベントを経験するユーザにとって有用な情報を含む可能性が高まると考え，他のトピックと共起した記事数を考慮したスコアリングを行う．最後に，2 つのスコアで上位となったトピックが高い割合で出現するブログ記事を提示する．

本論文の構成を以下に示す．2 章では，関連研究を紹介し，本研究の位置付けを述べる．3 章では，これからそのライフイベントを経験するユーザにとって有用な情報を含むブログ記事を提示するためのトピックを抽出する提案手法について，詳細を述べる．4 章では，「出産」，「就職」，「結婚」，「大学入学」，「子供の小学校入学」の 5 つのライフイベントに着目した実験データについて説明し，その妥当性について検証する．5 章では，4 章で得られたデータを用いて，提案手法により抽出したトピックが高い割合で出現するブログ記事を，各ライフイベントを経験したことのある 6 名の実験参加者に提示し，トピックに関連する記事数のバーストを考慮した手法と比較することで，提案手法の有効性を確認する．最後に，6 章で，本研究で得られた知見をまとめる．

[*8] 特定のユーザが 1 つのトピックについて多くの記事を投稿する影響を避けるために，ブログ記事数とはせず，投稿人数とする．

2. 関連研究

2.1 ライフイベントを経験したユーザの興味や行動の変化に関する研究

本節では，ライフイベントを経験したユーザの興味や行動の変化に着目した研究のうち，調査する興味や行動を事前に設定する研究と，興味や行動の変化を自動抽出する研究を紹介する．

2.1.1 事前に調査する興味や行動を設定する研究

Choudhury ら [5] は，Twitter 上で自身の「婚約」について投稿しているユーザを収集し，使用する単語や投稿内容の変化を分析した．分析の結果，“fiancé,” “fiancée,” “husband,” “wife” などの特定の単語の使用割合が増加することを明らかにした．さらに，婚約宣言の前後では，結婚式に関する投稿や，交際相手との交流に関する投稿が増加することを示した．また，Burke ら [4] は，「失職」を経験したユーザを Facebook 上の広告やメールで募集し，ストレスの変化や新たな職の獲得までの Facebook 上での活動について分析した．分析の結果，失職後の Facebook 上でのコミュニケーションは，ストレスの軽減や新たな職を見つける際に有効であることを示した．

これらの研究では，事前に調査する興味や行動を設定しているため，ライフイベント前後で変化する興味や行動が明らかでない場合に用いることができない．また，単純なクエリで，ブログ記事集合や SNS の投稿を検索した場合，大量のノイズが含まれるため，有用な情報が含まれる投稿を検索することは困難である．本研究では，これからライフイベントを経験するユーザの支援を目的とし，時系列トピックモデルにより推定したライフイベントを経験したユーザのブログ記事に出現するトピックから，有用な情報を含むブログ記事を提示するためのトピックを抽出する．

2.1.2 興味や行動の変化を自動抽出する研究

我々は，本研究と同様に，ライフイベントを経験したユーザの興味や行動は，投稿するブログ記事のトピックとして出現すると考え，「出産」，「就職」，「結婚」の 3 つのライフイベントにおいて，時系列トピックモデルを用いて，ブログ記事からトピックを推定し，ライフイベントを機に出現確率が大きく変化したトピックを選択するための手法を提案した [15]．これにより，ライフイベントを機に変化する興味や行動を自動抽出することができる．

この研究では，時系列トピックモデルで推定したトピック集合から，ライフイベントを機に変化したユーザの興味や行動を反映したトピックを抽出している．本研究では，ユーザの興味や行動の変化を反映しており，かつ，ライフイベントと関わりの深いトピックが，有用な情報を含むブログ記事に出現するトピックであると仮定し，これらの特徴を持つトピックを抽出する手法を提案する．

2.2　時系列トピックモデルを用いた研究

本研究では，ブログ記事に出現するトピック分布を推定するために，LDA（Latent Dirichlet Allocation）[3] に時系列情報を加えた拡張モデルである DTM [2] を用いる．本節では，トピックモデルを用いた時系列変化の分析を行っている研究を紹介する．

Zhang ら [18] は，Twitter 上での商品ブランドに関するツイートや画像に DTM を適用するためのモデルを提案し，商品ブランドの盛衰を分析した．また，Kim ら [8] は，任意の時期ごとに分割したデータから，PLSA [7] を用いてトピック推移を抽出し，株価の変動などの時系列データとの因果関係のあるトピック推移のみを抽出する手法を提案した．特定のトピックを抽出する手法としては，Wang ら [16] が，時期によりメディアとユーザの関心が変化するニューストピックについて，その時期にメディアの報道姿勢とユーザの関心の両方が高いニューストピックを上位とするランキング手法を提案している．

これらの研究では，様々なデータに時系列トピックモデルを適用し，時期ごとのユーザの興味の変化をトピックとして抽出している．本研究では，ライフイベントを経験したユーザの投稿したブログ記事に時系列トピックモデルを適用し，時期ごとのブログ記事に出現するトピック分布を推定する．さらに，これからライフイベントを経験するユーザに有用な情報を提示するために，推定したトピックについて，ユーザの興味や行動の変化を反映し，ライフイベントと関わりが深いトピックを抽出する手法を提案する．

2.3　時系列変化の特徴を利用したトピック抽出

本研究では，各トピックの時系列変化の特徴を利用して，ユーザの興味や行動の変化を反映したトピックを抽出する．本節では，特定のトピックを抽出するために，トピックの時系列変化の特徴を利用した研究を紹介する．

水田ら [23] は，LDA を利用したトピックの推定に，時間フィルタを組み合わせることで，バースト性のあるトピックを抽出できる t-LDA 法を提案した．また，Koike ら [10] は，時系列トピックモデルで推定したトピックに，Kleinberg のバースト検出手法 [9] を適用し，特定の時期にバーストしたトピックを抽出する手法を提案している．

これらの研究のように，ニュース記事や SNS を対象として，バーストしたトピックに着目する研究が多数報告されている [11], [22], [24]．一方で，ライフイベントを経験したユーザの投稿するブログ記事に出現するトピックの場合，「出産」イベントの「陣痛」トピックなどの特定の時期の体験を反映したバーストしやすいトピックだけでなく，「育児」トピックなどの，ライフイベント後に出現確率が増加し，その後一定の確率となる，ユーザの生活の変化を反映したトピックを抽出することも重要である．本研究では，これらの時系列変化の特徴を考慮し，トピックを抽出する．

2.4　トピックどうしの共起に着目したトピック抽出

本研究では，時系列変化の特徴だけでなく，ブログ記事におけるトピックどうしの共起に着目して，有用な情報を含むブログ記事に出現するトピックを抽出する．本節では，トピックどうしの共起に着目した研究を紹介する．

Zhou ら [19] は，論文の共著関係を利用して，研究トピックのトレンド分析を行っている．その際に，研究トピックは互いに依存しており，トピックどうしは共起しながら，別のトピックに変化していくと仮定している．また，Wang らの提案した時系列トピックモデル TOT（Topics Over Time）[17] は，各トピックについて，時系列変化をベータ分布に基づいて推定する手法であり，時期ごとのトピックどうしの共起を明らかにすることができる．

これらの研究では，トピックのトレンドが時期ごとに共起しつつ，変化していくことを仮定している．本研究では，ライフイベントを経験したユーザの興味や行動の変化を反映したトピックのうち，同一の時期に共通するトピックをブログ記事に共起させて書かれることが相対的に多いトピックを，ライフイベント経験に関わりの深いトピックと考え，有用なブログ記事に出現するトピックと仮定する．また，他のほとんどのトピックと共起するトピックはノイズとしてフィルタリングする．

3.　ライフイベントを経験するユーザにとって有用なトピックの抽出手法

本研究では，ライフイベントを経験したユーザが投稿したブログ記事を分析することで，これからライフイベントを経験するユーザにとって有用な情報を含むブログ記事を提示するためのトピックを抽出する手法を提案する．**図 1** に，提案手法の一連の流れを示す．

まず，3.1 節で，DTM [2] を用いた，ライフイベントを経験したユーザ集合のブログ記事に出現するトピック分布の推定について説明する．次に，3.2 節で，ブログ記事内において，ほとんどのトピックと共起するトピックをノイズとしてフィルタリングする手法について説明する．さらに，3.3 節で，各トピックについて，ライフイベントの前後および特定の時期における投稿人数の増加と，ブログ記事内における他のトピックとの共起とを考慮したスコアを計算し，上位のトピックを抽出する手法について説明する．最後に，3.4 節で，得られたトピックが高い割合で出現するブログ記事を提示する手法について説明する．

3.1　ブログ記事に対する時系列を考慮したトピック分布の推定

本研究では，ブログ記事におけるトピック分布の推定に，LDA に時系列情報を加えた拡張モデルである DTM を用いる．DTM を用いることで，同一のトピックの時間発展を追跡することが可能となる．

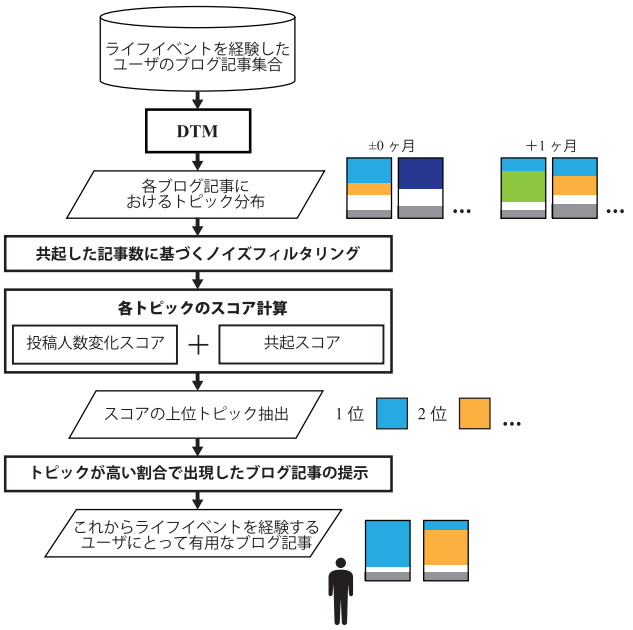


図 1　有用な情報を含むブログ記事に出現するトピックの抽出
Fig. 1　Topic extraction in blog posts including some useful information.

まず，ライフイベントを経験したユーザ集合によって書かれたブログ記事について，ライフイベントを経験した月を ±0 カ月とし，経過月ごと[*9]に集約する．次に，経過月を単位としたブログ記事の集合をそれぞれ形態素単位に分割し，DTM により，各ブログ記事におけるトピック分布を推定する．なお，DTM では，時間分割数パラメータ TS により，各トピックの月ごとの確率分布と，単語の確率分布が TS 個生成される．今回の分析では，ライフイベント以前の 12 カ月間，ライフイベントを経験した月，ライフイベント以後の 12 カ月間の計 25 カ月間を分析対象とするため，$TS = 25$ とした．

3.2　共起した記事数を用いたノイズフィルタリング

本節では，各トピックと他のトピックとのブログ記事における共起に着目したノイズフィルタリングについて説明する．

前節の処理で得られたトピック集合には，「思う，言う，感じ，気持ち」などの自身の感情に関する単語が多く含まれたトピック（「感情」トピック）や，「今日，明日，時間，予定，起きる，帰る」などの習慣的にブログ記事に出現する単語が多く含まれたトピック（「習慣」トピック）が存在する．ライフイベントを経験することで，環境や習慣が変わることから，ライフイベント後に「感情」トピックや「習慣」トピックの出現確率が増加する場合がある．しかし，このようなトピックが高い割合で出現するブログ記事には，短文が多く，これからライフイベントを経験するユーザにとって，有用な情報が含まれていることは少ない．

[*9] ユーザのライフイベント経験の時間軸を合わせるために，経過月ごとにブログ記事を分割する．また，日単位や週単位では，十分なブログ記事数が得られないため，分析は月単位で行った．

表 2　「感情」トピックのブログ記事例
Table 2　Example of blog post for “emotional” topic.

ライフイベント	ブログ記事
「就職」	最近涙脆くて大変です．最近すぐに泣けてきます．どうして私はこんなにも泣き虫なんでしょう．強くなりたいです．

表 2 に，「就職」イベントにおける「感情」トピックが高い割合で出現するブログ記事の例を示す．

本研究では，このようなトピックをノイズと見なす．また，このようなトピックは，他の多くの異なるトピックと共起することから，ブログ記事におけるトピックどうしの共起を計算することで事前にフィルタリングする．たとえば，「感情」トピックの「思う，言う，感じ」などの単語は，他の主題となるトピックの体験に関する感想として出現することが多い．また，「習慣」トピックの「今日，明日，時間」などの単語は，他の主題となるトピックの単語と同時に出現することが多い．これらの特徴を考慮し，本研究では，まず，各トピックについて，他の全トピックとの共起した記事数の総和を計算する．全トピック数が K のとき，トピック t_i 以外の全トピックとの共起した記事数の総和 $n(t_i)$ は以下の式で定義する．

$$n(t_i) = \sum_{j:j\neq i}^{K} |d(t_i) \cap d(t_j)| \tag{1}$$

ここで，$d(t_i)$ は，トピック t_i が出現したブログ記事集合を示す．なお，本研究では，トピックの出現確率が 0.3 以上のとき，そのトピックがブログ記事に出現したと判断する．

次に，$n(t_i)$ の値が有意に大きいトピック t_i を外れ値としてフィルタリングする．外れ値の検出には，3σ 法を用いる．3σ 法は，あるデータの偏差が母集団の標準偏差の 3 倍より大きいときに，外れ値とする手法で，複数の研究で用いられている [14], [24]．本研究では，トピック t_i 以外の全トピックとの共起した記事数の総和 $n(t_i)$ の値について，次式を満たす場合に，トピック t_i をノイズトピックとしてフィルタリングする．

$$n(t_i) > \mu_n + 3\sigma_n \tag{2}$$

ここで，μ_n は，全トピックの n の値の平均値，σ_n は，全トピックの n の値の標準偏差を示す．

3.3　投稿人数変化スコアと共起スコアによるトピック抽出

本節では，ライフイベントの前後および特定の時期における投稿人数の増加に着目したトピックの投稿人数変化スコアと，ブログ記事内における他のトピックとの共起を考慮した共起スコアの計算について説明する．

3.3.1　トピックの投稿人数変化スコア

本研究では，ライフイベントを経験したユーザの興味や行動の変化を反映したトピックを抽出するために，各トピックについて，投稿人数の変化に着目した投稿人数変化スコアを計算する．別の手法として，トピックの出現確率の推移を利用する手法 [15] や，時期ごとのブログ記事における，トピックの出現確率の和の変化を利用する手法 [10] が考えられるが，個人のブログ記事に適用する場合，1 人のユーザが同じトピックに関するブログ記事を何度も投稿するとバイアスが生じるため，投稿人数の変化を利用する．

各月における投稿人数のベクトル化：まず，各トピックについて，分析期間である 25 カ月間の各月における投稿人数を計算し，25 次元のベクトルを作成する．各トピックにおける月ごとの投稿人数は，トピックの出現確率が 0.3 以上となるブログ記事をその月に投稿した人数とする．また，各月ごとのユーザ数は一定ではないため，その月の全投稿人数で，ベクトルを正規化する．以下に，トピック t_i に関するある月 m の投稿人数 $u(t_i, m)$ を正規化するための式を示す．

$$u(t_i, m)_{normalized} = \frac{u(t_i, m)}{u(m)} \tag{3}$$

ここで，$u(m)$ は，その月の全投稿人数を示す．

投稿人数変化スコアの計算：次に，ライフイベントを経験したユーザの興味や行動の変化を反映したトピックを抽出するために，正規化した 25 次元のベクトルについて，ライフイベントの前後および特定の時期におけるトピックに関する投稿人数の増加に着目した投稿人数変化スコアを計算する．なお，本研究では，ライフイベント経験からの経過月が −12 カ月から ±0 カ月の 13 カ月間の投稿をライフイベント前の投稿，±0 カ月から +12 カ月までの 13 カ月間の投稿をライフイベント後の投稿と見なす．

図 2 に，投稿人数変化スコアでとらえる特徴を示す．図 2 の赤線のグラフは，ライフイベントの以前だけ，あるいは以後だけに着目すると，投稿人数の変化の幅は大きくないが，ライフイベント以後には投稿人数が増加している．この特徴は，「出産」イベントにおける「育児」トピックや，「就職」イベントにおける「会社生活」トピックなどの，ライフイベントの影響で日常的に行うようになった生活の変化を反映したトピックに現れる．提案手法では，このような特徴を持つトピックを抽出するために，ライフイベント前後における投稿人数の平均の差を計算する．一方，図 2 の青線のグラフは，ライフイベント直後の ±0 カ月に投稿人数が急増し，ライフイベント後に減少している．この特徴は，「出産」イベントにおける「陣痛」トピックや，「大学入学」イベントにおける「テスト・レポート」トピックなどの，ライフイベントを経験するユーザの多くが特定の時期に体験する興味や行動を反映したトピックに現れる．提案手法では，このような特徴を持つトピックを抽出するた

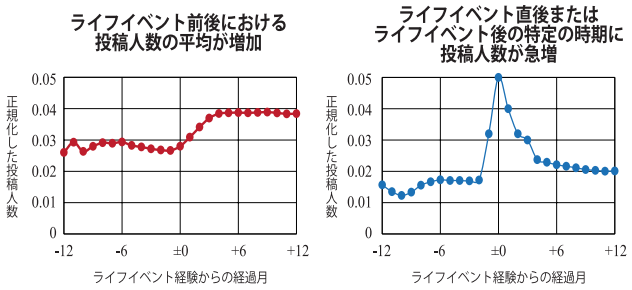


図 2　投稿人数変化スコアでとらえる特徴

Fig. 2　Trends detected by transition score for number of posting users.

めに，ライフイベント後の月[*10]における最大の投稿人数とその他の月の投稿人数の平均との差を計算する．以下で，これらの特徴を考慮した投稿人数変化スコアの計算方法について説明する．

まず，それぞれのトピックについて，ライフイベント前後における投稿人数の平均の差を計算し，Z スコアで正規化する．トピック t_i の投稿人数の平均の差の Z スコア $m\text{-}score_{t_i}$ を計算するための式を以下に示す．

$$m\text{-}score_{t_i} = \frac{(E(\boldsymbol{y}_{t_i}) - E(\boldsymbol{x}_{t_i})) - \mu_m}{\sigma_m} \tag{4}$$

$$\mu_m = \frac{1}{K}\sum_{j}^{K}(E(\boldsymbol{y}_{t_j}) - E(\boldsymbol{x}_{t_j})) \tag{5}$$

$$\sigma_m = \sqrt{\frac{1}{K}\sum_{j}^{K}((E(\boldsymbol{y}_{t_j}) - E(\boldsymbol{x}_{t_j})) - \mu_m)^2} \tag{6}$$

ここで，ライフイベントを経験する前のすべての月に対する，式 (3) で正規化した投稿人数を要素としたベクトルは $\boldsymbol{x}_{t_i}$，経験した後のベクトルは $\boldsymbol{y}_{t_i}$，K は全トピック数，$E(\boldsymbol{x}_{t_i})$ は $\boldsymbol{x}_{t_i}$ の平均，μ_m は全トピックにおける，ライフイベント前後における投稿人数の平均の差の平均値，σ_m は全トピックにおける，ライフイベント前後における投稿人数の平均の差の標準偏差をそれぞれ表す．

次に，それぞれのトピックについて，ライフイベント後の月における最大の投稿人数とその他の月の投稿人数の平均との差を計算し，Z スコアで正規化する．トピック t_i の最大の投稿人数とその他の月の投稿人数の平均との差の Z スコア $r\text{-}score_{t_i}$ を計算するための式を以下に示す．

$$r\text{-}score_{t_i} = \frac{(\max(\boldsymbol{y}_{t_i}) - E(\boldsymbol{u}_{t_i} \setminus \arg\max(\boldsymbol{y}_{t_i})) - \mu_r}{\sigma_r} \tag{7}$$

$$\mu_r = \frac{1}{K}\sum_{j}^{K}(\max(\boldsymbol{y}_{t_j}) - E(\boldsymbol{u}_{t_j} \setminus \arg\max(\boldsymbol{y}_{t_j})) \tag{8}$$

$$\sigma_r = \sqrt{\frac{1}{K}\sum_{j}^{K}(\max(\boldsymbol{y}_{t_j}) - E(\boldsymbol{u}_{t_j} \setminus \arg\max(\boldsymbol{y}_{t_j}) - \mu_r)^2} \tag{9}$$

[*10] ライフイベントを経験した後にユーザが体験する興味や行動を抽出するため．

ここで，トピック t_i のすべての月に対する，式 (3) で正規化した投稿人数を要素としたベクトルは $\boldsymbol{u}_{t_i}$，ライフイベントを経験した後の投稿人数のベクトルは $\boldsymbol{y}_{t_i}$，K は全トピック数，$\max(\boldsymbol{y}_{t_i})$ は，$\boldsymbol{y}_{t_i}$ の最大の投稿人数，$\arg\max(\boldsymbol{y}_{t_i})$ は，最大の投稿人数となった月の要素，$(\boldsymbol{u}_{t_i} \setminus \arg\max(\boldsymbol{y}_{t_i}))$ は，正規化した投稿人数を要素としたベクトルから，最大の投稿人数となった月の要素を除いたベクトル，$E(\boldsymbol{u}_{t_i} \setminus \arg\max(\boldsymbol{y}_{t_i}))$ は，正規化した投稿人数を要素としたベクトルから，最大の投稿人数となった月の要素を除いたベクトルの平均，μ_r は全トピックにおける，最大の投稿人数とそのほかの月の投稿人数の平均との差の平均値，σ_r は全トピックにおける，最大の投稿人数とその他の月の投稿人数の平均との差の標準偏差をそれぞれ表す．

最後に，トピック t_i の投稿人数変化スコア $T\text{-}score_{t_i}$ は，$m\text{-}score_{t_i}$ と $r\text{-}score_{t_i}$ の高い方の値とする．これによりライフイベントにおいて，どちらかの特徴を強く反映したトピックを抽出することが可能となる．以下に，トピック t_i の投稿人数変化スコア $T\text{-}score_{t_i}$ の式を示す．

$$T\text{-}score_{t_i} = \max(m\text{-}score_{t_i}, r\text{-}score_{t_i}) \tag{10}$$

3.3.2　トピックの共起スコア

本項では，他のトピックとの共起のしやすさを示すトピックの共起スコアの計算について説明する．「就職」を経験したユーザのブログ記事におけるトピック分布を**図 3** に示す．図 3 に示すように，「就職」を経験したユーザのブログ記事内のトピック分布では，「会社生活」トピックと「残業」トピックが共起しやすい[*11]．本研究では，これらのようなライフイベントと関わりが深いトピックどうしは，同じブログ記事内で共起しやすく，ユーザの興味や行動を反映する，と仮定する．投稿人数変化スコアと，トピックどうしの共起を考慮した共起スコアとを利用することで，ユーザの興味や行動の変化を反映し，ライフイベントと関わりが深いトピックを抽出することができる．また，投稿人数変化の特徴だけでなく，ブログ記事内の共起を考慮することで，「クリスマス」トピックなどの特定の時期に投稿人数が急増する季節性のトピックや，「大学入学」の「アニメ・漫画」トピックなどのライフイベントの影響で間接的に投稿する人数が増加したと思われるトピックを除外することが可能となる[*12]．なお，共起スコアを計算するうえで，「感情」トピックや「習慣」トピックのようなつねにどのトピックとも共起するノイズトピックは 3.2 節で説明したように，事前にフィルタリングする．

共起スコアは，ノイズトピックを除いたトピック集合において，他の全トピックとの共起した記事数を，Z スコアで正規化したものと定義する．トピック t_i の共起スコア

*11 ほかに，「出産」イベントにおける「育児」トピックと「乳児の成長」トピックなど．

*12 これらのトピックは，ライフイベントと直接的な関わりがないため，単独で現れることが多い．

図 3　「就職」を経験したユーザのブログ記事におけるトピック分布
Fig. 3　Topic distribution in blogs posted by users “finding employment”.

$C\text{-}score_{t_i}$ を計算するための式を以下に示す．

$$C - score_{t_i} = \frac{\sum_{j:j\neq i}^{L} |d(t_i) \cap d(t_j)| - \mu_c}{\sigma_c} \tag{11}$$

$$\mu_c = \frac{1}{L} \sum_{j}^{L} \sum_{k:k\neq j}^{L} |d(t_j) \cap d(t_k)| \tag{12}$$

$$\sigma_c = \sqrt{\frac{1}{L} \sum_{j}^{L} \left(\sum_{k:k\neq j}^{L} |d(t_j) \cap d(t_k)| - \mu_c \right)^2} \tag{13}$$

ここで，L はノイズを除いた全トピック数，$\sum_{j:j\neq i}^{L} |d(t_i) \cap d(t_j)|$ は，トピック t_i と，L から t_i を除いたトピック集合におけるそれぞれのトピックとが共起した記事数の合計，μ_c はノイズを除いた全トピックにおける，他のトピックと共起した記事数の合計の平均値，σ_c はノイズを除いた全トピックにおける，他のトピックと共起した記事数の標準偏差を，それぞれ表す．なお，本研究では，トピックの出現確率が 0.3 以上のとき，そのトピックがブログ記事に出現したと判断する．

本研究では，これまでに説明した投稿人数変化スコアと共起スコアの和が上位となったトピックを，ライフイベントを経験したユーザの興味や行動の変化を反映し，ユーザの体験に関する有用な情報を含むブログ記事に出現するトピックとする．以下に，トピック t_i のスコアを計算する式を示す．

$$score_{t_i} = T\text{-}score_{t_i} + C\text{-}score_{t_i} \tag{14}$$

3.4　有用な情報を含むブログ記事の提示

本研究では，3.3 節で抽出したトピックが高い割合で出現したブログ記事を提示する．ただし，提示するブログ記事が短文の場合，有用な情報が含まれにくいと考え，文字数が 500 文字以上のブログ記事について，抽出したトピックの割合が高い順に提示する．

4.　ライフイベントを経験したユーザのブログ記事集合の構築

本章では，実験データについて説明する．今回の分析では，多くのユーザがブログ記事や SNS 上で報告する「出産」，「就職」，「結婚」[*13]，「大学入学」，「子供の小学校入学」

*13 結婚式をあげたことの報告だけでなく，入籍の報告も「結婚」と判断する．

表 3　ブログ記事を抽出するためのクエリ一覧
Table 3　Queries used for retrieval of blog posts.

ライフイベント	クエリ
「出産」	「出産しました」
「就職」	「新社会人」OR「入社式」
「結婚」	「入籍しました」OR「結婚しました」
「大学入学」	「大学」AND「入学式でした」
「子供の小学校入学」	「小学校」AND「入学式でした」

の 5 つのライフイベントを対象として実験を行う．これらのライフイベントは，ユーザにとってそれまでしたことのない行動をするようになることが多く，有用な情報を求めると考え，選択した．実験データの抽出対象は，ブログランキングサイトである blogram.jp[*14]に登録されているブログ記事のうち，2008 年 1 月 11 日から 2011 年 3 月 13 日までのブログ記事とした．なお，「出産」，「就職」，「結婚」の実験データは文献 [15] と同一のものである．

4.1　ライフイベントを経験したユーザの選択

ライフイベントを経験したユーザ集合を選択するために，ライフイベントに関連するクエリを含むブログ記事を抽出し，人手でライフイベント経験の有無をラベリングする．各ライフイベントで利用するクエリの一覧を**表 3** に示す．

各ライフイベントについて，これらのクエリに該当するブログ記事を投稿しているユーザ集合の全ブログ記事を抽出した．しかし，このようにしてブログ記事を抽出した場合，実際に該当のライフイベントを経験しているユーザのブログ記事だけでなく，ユーザの未来の予定や過去の出来事を回想している記事，ユーザが体験の主体ではない記事，宣伝用のブログ記事などのノイズが含まれる [21]．そのため，著者が「ライフイベントを経験している」と判断したユーザを，各ライフイベントごとに 100 名ずつ抽出した．この際，投稿しているブログ記事が少なすぎるユーザは，分析の際に，ライフイベントに関連する興味や行動が十分に得られない恐れがあるため，30 件以上のブログ記事を投稿しているユーザのみを抽出した．

「出産」を対象としたラベリングの結果とブログ記事の一部の例を**表 4** に示す．ここで，ラベル番号 1 が「該当のライフイベントを経験している」と判断されたブログ記事，ラベル番号 0 が「該当のライフイベントを経験していない」と判断されたブログ記事である．

このようにして得られた実験データは，各ライフイベントごとに 100 ユーザが投稿した「出産」34,753 記事，「就職」40,238 記事，「結婚」30,605 記事，「大学入学」28,195 記事，「子供の小学校入学」31,073 記事となった．それぞれのライフイベントの投稿時期におけるブログ記事数とユーザ数を**表 5**，**表 6**，**表 7**，**表 8**，**表 9** に示す．

[*14] http://blogram.jp/

表 4　「出産」のラベリング結果の例
Table 4　Examples of labels for “childbirth” experiences.

ラベル	ブログ記事の一部
1	39 週 1 日，今日の明け方に 3100 g の元気な男の子を出産しました．
1	予定日より 10 日遅れで出産しました．お産は大変だったけど，母子共に無事に退院でき本当に良かったです．
0	私の親友が 2 人目の男の子を出産しました．

表 5　「出産」における投稿時期ごとのブログ記事数とユーザ数
Table 5　Numbers of blog posts and users in “childbirth” life event by posting months.

投稿時期	記事数	ユーザ数	投稿時期	記事数	ユーザ数
−12 カ月	909	36	±0 カ月	1,734	100
−11 カ月	1,064	41	+1 カ月	1,179	100
−10 カ月	1,139	47	+2 カ月	1,396	100
−9 カ月	1,277	51	+3 カ月	1,541	97
−8 カ月	1,325	61	+4 カ月	1,525	92
−7 カ月	809	64	+5 カ月	1,526	90
−6 カ月	932	69	+6 カ月	1,545	91
−5 カ月	1,169	75	+7 カ月	1,495	92
−4 カ月	1,395	81	+8 カ月	1,621	87
−3 カ月	1,607	88	+9 カ月	1,557	89
−2 カ月	1,727	95	+10 カ月	1,434	86
−1 カ月	1,941	99	+11 カ月	1,481	85
			+12 カ月	1,425	85

表 6　「就職」における投稿時期ごとのブログ記事数とユーザ数
Table 6　Numbers of blog posts and users in “finding employment” life event by posting months.

投稿時期	記事数	ユーザ数	投稿時期	記事数	ユーザ数
−12 カ月	993	26	±0 カ月	2,757	100
−11 カ月	1,050	27	+1 カ月	2,599	95
−10 カ月	1,042	30	+2 カ月	1,863	93
−9 カ月	1,024	33	+3 カ月	1,810	91
−8 カ月	1,334	35	+4 カ月	1,747	90
−7 カ月	1,548	43	+5 カ月	1,596	88
−6 カ月	1,776	47	+6 カ月	1,538	86
−5 カ月	1,675	49	+7 カ月	1,387	83
−4 カ月	1,703	50	+8 カ月	1,512	83
−3 カ月	1,583	58	+9 カ月	1,526	84
−2 カ月	1,922	89	+10 カ月	1,308	84
−1 カ月	2,830	100	+11 カ月	1,096	87
			+12 カ月	1,019	64

4.2　ラベリングの妥当性の検証

著者によるラベリングの妥当性を検証するために，著者が「ライフイベントを経験している」と判断した 50 名のユーザと「ライフイベントを経験していない」と判断した 50 名のユーザを各ライフイベントごとに抽出した．このようにして得られた各ライフイベントごとの 100 ユーザについて，表 3 のクエリが現れたブログ記事を選択し，アノテータに判定させた．ただし，「ライフイベントを経験している」と判断したユーザについては，著者が 4.1 節で，

表 7　「結婚」における投稿時期ごとのブログ記事数とユーザ数
Table 7　Numbers of blog posts and users in “marriage” life event by posting months.

投稿時期	記事数	ユーザ数	投稿時期	記事数	ユーザ数
−12 カ月	788	39	±0 カ月	1,737	100
−11 カ月	806	47	+1 カ月	1,566	100
−10 カ月	824	50	+2 カ月	1,464	92
−9 カ月	901	53	+3 カ月	1,424	91
−8 カ月	1,084	58	+4 カ月	1,192	86
−7 カ月	1,136	60	+5 カ月	1,187	84
−6 カ月	1,219	66	+6 カ月	1,085	83
−5 カ月	1,477	74	+7 カ月	1,036	79
−4 カ月	1,614	78	+8 カ月	936	75
−3 カ月	1,622	88	+9 カ月	955	74
−2 カ月	1,727	95	+10 カ月	970	71
−1 カ月	1,686	98	+11 カ月	1,001	72
			+12 カ月	1,168	71

表 8　「大学入学」における投稿時期ごとのブログ記事数とユーザ数
Table 8　Numbers of blog posts and users in “university admission” life event by posting months.

投稿時期	記事数	ユーザ数	投稿時期	記事数	ユーザ数
−12 カ月	588	33	±0 カ月	2,144	100
−11 カ月	676	33	+1 カ月	1,951	100
−10 カ月	551	34	+2 カ月	1,673	100
−9 カ月	682	40	+3 カ月	1,544	100
−8 カ月	899	42	+4 カ月	1,402	96
−7 カ月	829	44	+5 カ月	1,356	94
−6 カ月	838	46	+6 カ月	1,327	94
−5 カ月	785	50	+7 カ月	1,131	90
−4 カ月	866	55	+8 カ月	1,108	85
−3 カ月	781	61	+9 カ月	1,064	83
−2 カ月	1,380	84	+10 カ月	991	83
−1 カ月	1,784	98	+11 カ月	998	84
			+12 カ月	847	67

表 9　「子供の小学校入学」における投稿時期ごとのブログ記事数とユーザ数
Table 9　Numbers of blog posts and users in “child’s admission to elementary school” life event by posting months.

投稿時期	記事数	ユーザ数	投稿時期	記事数	ユーザ数
−12 カ月	703	33	±0 カ月	1,874	100
−11 カ月	683	34	+1 カ月	1,776	99
−10 カ月	638	34	+2 カ月	1,638	98
−9 カ月	740	39	+3 カ月	1,692	96
−8 カ月	812	43	+4 カ月	1,586	96
−7 カ月	876	53	+5 カ月	1,545	94
−6 カ月	1,116	56	+6 カ月	1,501	90
−5 カ月	1,017	59	+7 カ月	1,381	90
−4 カ月	1,070	63	+8 カ月	1,528	87
−3 カ月	1,073	73	+9 カ月	1,459	90
−2 カ月	1,399	91	+10 カ月	1,234	86
−1 カ月	1,575	99	+11 カ月	1,148	86
			+12 カ月	1,009	67

「該当のライフイベントを経験している」と判断した記事を用いた．アノテータは，20 代男性の大学生 2 名である．アノテータは，ユーザがその記事を投稿した時間かその前後で，「該当のライフイベントを経験しているか否か」を判断し，ラベリングする．ラベリングの結果，今回提示した 5 つのライフイベントにおけるラベリングの判定者間一致率は 100%であった[*15]．この結果から，著者によるラベリングは信頼性が高いと考え，4.1 節で著者が判断した各ライフイベントごとに 100 ユーザが投稿したブログ記事を実験データとする．

5. 実験：有用な情報を含むブログ記事に出現するトピックの抽出

5.1 実験方法

本研究では，「出産」，「就職」，「結婚」，「大学入学」，「子供の小学校入学」の 5 つのライフイベントを対象として，提案手法および比較手法により，有用なトピックを抽出する．また，抽出したトピックが出現するブログ記事が，これからライフイベントを経験するユーザにとって有用であるか否かについて，評価する．まず，提案手法と比較手法を用いて，各ライフイベントにおいて，上位 5 件のトピックを抽出する．次に，3.4 節で説明した手法で，それぞれのトピックが高い割合で出現するブログ記事を，各トピックについて 3 件ずつ実験参加者に提示する．

「出産」，「結婚」，「子供の小学校入学」を評価する実験参加者は，出産経験があり育児中の 3 名の主婦，「就職」，「大学入学」を評価する実験参加者は，4 年制大学を卒業後，新卒として社会人を 1 年以上経験した 3 名（男性 1 名，女性 2 名）とした．実験参加者は，そのライフイベントを経験する時期にブログ記事を読んだと仮定して，ブログ記事に含まれる情報の有用性を評価する．具体的には，ブログ記事中に**自分がそのライフイベントを経験する上で，知っておいた方が良かった情報が含まれるか**否かの 2 値で評価する．まず，各ブログ記事について，実験参加者の 3 名中 2 名以上が，自分がそのライフイベントを経験する上で，知っておいた方が良かった情報が含まれる，と判断したブログ記事を有用なブログ記事とする．次に，有用なブログ記事を 1 件以上含むトピックを有用なトピックとする．最後に，抽出した 5 件のトピックに含まれる有用なトピックの割合で評価する．

5.2 比較手法

2.3 節で述べたように，ニュース記事や SNS を対象として，バーストしたトピックに着目する研究が報告されている．また，研究論文のトレンド分析のために，出現する単語ペアのバーストに着目した研究が報告されている [20]．ライフイベントを経験したユーザのブログ記事集合において，バーストするトピックは，同一の時期に多くのユーザ

[*15] 表 4 に示すような明確にそのライフイベントを経験していると判断できるブログ記事を対象としているため，完全に一致したと考えられる．

がそのトピックを集中的に投稿していることを表すため，その時期に起きやすい体験を反映したトピックとなる．このようなトピックは，たとえば，「出産」イベントにおける「陣痛」トピックや，「入院準備」トピックが該当し，ユーザの体験に基づく情報が含まれたブログ記事を提示することができる．

そこで，比較手法として既存のバースト検出手法である，Kleinberg の手法 [9] を用いる．Kleinberg のバースト検出手法は，離散時間で到着する文書の状態を，通常状態とバースト状態の 2 つの状態からなるオートマトンでモデル化することで，各時刻における状態を推定する手法である．ある単語のバーストをとらえる場合，時期ごとの全文書数におけるその単語が出現した文書数の割合によって，状態を推定する．さらに，バースト状態となった区間における，バーストの度合い（バースト度）を計算できる．

本研究では，Kleinberg のバースト検出手法を用いて，あるトピックのバーストをとらえる際に，そのトピックが 0.3 以上の割合で出現したブログ記事数と，ライフイベントからの経過月ごとの全ブログ記事数を利用する[*16]．また，ライフイベント後にバーストが検出されたトピックについて，バースト度を計算し，上位 5 件のトピックを比較対象として抽出した．なお，Kleinberg のバースト検出手法におけるモデルパラメータ s，γ については，予備実験により，すべてのライフイベントで $s = 2.0$，$\gamma = 1.0$ とした．

5.3　実験環境

DTM の実装には，Python のライブラリである gensim [*17] を用いた．トピック数は，各トピックの独立性を評価することで決定した．まず，トピック数を 10 から 100 までの 10 刻みで変動させ，出力された各トピックの確率分布間の非類似度（dissimilarity）を JS-divergence により計算し，分析期間内における全トピックの組合せの平均が最も高いものとした．確率分布 P，Q 間の JS-divergence は，以下の式で計算する．なお，M は P と Q の平均であり，$M(i) = \frac{P(i)+Q(i)}{2}$ である．

$$JSD(P \parallel Q) = \frac{1}{2}\left(\sum_i P(i)\log\frac{P(i)}{M(i)}\right) + \frac{1}{2}\left(\sum_i Q(i)\log\frac{Q(i)}{M(i)}\right) \quad (15)$$

上記の式に基づき，出力トピック数は，「出産」で $K = 100$，「就職」で $K = 60$，「結婚」で $K = 80$，「大学入学」で $K = 90$，「子供の小学校入学」で $K = 90$ とした．なお，ハイパーパラメータは，すべてのライフイベントで $\alpha = 0.01$ とした．時間分割数は，3.1 節で述べたように $TS = 25$ とした．

[*16] 投稿人数の時期ごとの変化では，十分な数のトピックが検出されないライフイベントが複数あったため，記事数を用いた．
[*17] https://radimrehurek.com/gensim/

表 10　有用なトピックの割合
Table 10　Rates of useful topics.

ライフイベント	提案手法	比較手法
「出産」	**0.8**	0.6
「就職」	**0.6**	0.2
「結婚」	**0.4**	0.0
「大学入学」	**0.8**	0.6
「子供の小学校入学」	**0.6**	0.0
平均	**0.64***	0.28

* t-検定（片側検定，有意水準 1%，p=0.004）で有意に向上．

ブログ記事の形態素解析には MeCab [*18] を用いた．形態素解析の辞書は，mecab-ipadic-NEologd [*19] を利用した．分析する単語の品詞は名詞，動詞，形容詞とし，ひらがなのみ，または英数字のみで構成された 2 文字以下の単語は，ストップワードとして除外している．このようにして得られた総単語数は，「出産」で 7,938 語，「就職」で 15,561 語，「結婚」で 38,140 語，「大学入学」で 6,218 語，「子供の小学校入学」で 32,501 語で，異なり語数は，「出産」で 2,940 語，「就職」で 5,340 語，「結婚」で 5,624 語，「大学入学」で 2,263 語，「子供の小学校入学」で 7,134 語であった．

5.4　結果

提案手法と比較手法で抽出した，それぞれのライフイベントにおける有用なトピックの割合を**表 10** に示す．提案手法で抽出した有用なトピックの割合が，すべてのライフイベントで比較手法を上回ることを確認した．また，提案手法の評価結果の平均値が，比較手法を有意に上回ることを確認するために，有意水準 1%で対応のある片側 t 検定を適用したところ，$p = 0.004$ で有意差を確認できた．

提案手法と比較手法で抽出したトピックを**表 11** に示す．表中の下線のあるトピックが，実験参加者によって有用と判断されたトピックである．なお，表中のトピック名は，著者がトピックの単語分布と出現したブログ記事から判断しており，明らかに単語やブログ記事にまとまりがないと判断したトピックは省略している．また，有用な情報を含むと判断されたブログ記事の例を**表 12** に示す．提案手法を用いることで，表 12 に示すようなブログ記事をトピックごとに区別して提示することが可能となる．したがって，これからライフイベントを経験するユーザは，自身の興味に従って，トピックとブログ記事を選択できる．

5.5　考察

表 11 に示すように，提案手法では，すべてのライフイベントで 2 つ以上の有用なトピックを抽出できた．また，提案手法で抽出されたすべてのトピックは，単語にまとま

[*18] http://taku910.github.io/mecab/
[*19] https://github.com/neologd/mecab-ipadic-neologd

表 11　提案手法と比較手法で抽出したトピック
Table 11　Topics detected by proposed method and by comparison method.

ライフイベント	手法	トピック名
「出産」	提案手法	「育児」，「乳児の成長」，「乳児の風邪」，「陣痛」，「出産報告」
	比較手法	「入院準備」，「習慣」，「陣痛」，「胎児の様子」，「出産報告」
「就職」	提案手法	「残業」，「一人暮らしの料理」，「会社生活」，「感謝」，「大学の卒業」
	比較手法	「会社生活」，「感情」，「大学の卒業」
「結婚」	提案手法	「結婚式の準備」，「結婚相手の条件」，「結婚報告」，「新居での生活」，「妊娠準備」
	比較手法	「結婚報告」，「年賀状」
「大学入学」	提案手法	「大学受験」，「大学の授業」，「テスト・レポート」，「大学生活」，「食事」
	比較手法	「大学の授業」，「テスト・レポート」，「大学生活」，「旅行」
「子供の小学校入学」	提案手法	「子供の教材」，「小学校の授業」，「小学校生活」，「子供の習い事」，「食事」
	比較手法	「子供の夏休み」

表 12　提案手法で抽出できた有用と判断されたブログ記事の例
Table 12　Examples of blog posts evaluated useful in proposed method.

ライフイベント	トピック名	ブログ記事の一部（元の記事から抜粋した上で表現を一部編集）
「出産」	「育児」	... 今日で息子は 6 ヶ月になりました．・両方の手で上手におもちゃを持つことができるようになりました．片手ずつ，違うおもちゃを持てます．・支えがなくても，お座りをして左右に倒れなくなりました（前後には倒れる）．・歯茎がかゆいのか，舌でよく歯茎を舐めます．・4 ヶ月目より 5 ヶ月目の方が甘えん坊になった気がします．・誰にでもニコニコするのんびりでおっとりした性格です．周りの赤ちゃんたちもそれぞれの個性が強くでてきました．...
「就職」	「残業」	... 今月はあと 3 日ですが，残業時間が 40 時間になり，上司から帰れと言われました．計算したら，今月の給料は入社 5 年目の人に匹敵しそうです．ついでに会社の生命・損害保険に加入しとこう．月額 3,000 円で充実した保険内容だしね．さすが親元が大企業なだけある．...
「結婚」	「結婚式の準備」	... 招待状のあて名書きは，すべて自分たちで筆で書くようにプランナーさんに勧められました．その方が気持ちが伝わるから，と．招待状は 1 枚のカードになっていて，返信用はがき，式場までの地図と一緒に同封して完成です！...
「大学入学」	「大学受験」	... 私の目指す○○大学の△△学科ですが，受験科目は英語・数学・物理だけなのに，入学後は，高校レベルの化学はすべて知ってるものとして授業が進むようです．なので，留年率も非常に高いとのこと．その分，就職活動では楽できるそうです．...
「子供の小学校入学」	「子供の教材」	... ××という算数ドリルのシリーズは，他のものと違い，かわいい絵が多く，兄弟そろって，毎日進めることができています．...

りのあるトピックであった．これは，共起スコアを考慮することで，単独で出現する意味のないトピックを除外できたためと考えられる．

以下で，提案手法で用いた，投稿人数変化スコア，共起スコア，ノイズフィルタリングの各要素の精度への影響，提案手法で抽出できた有用なトピック，提案手法では抽出できなかった有用なトピックについて考察する．

5.5.1　提案手法で用いた各要素の精度への影響

提案手法で用いた，投稿人数変化スコア，共起スコア，ノイズフィルタリングの各要素の精度への影響を明らかにするために，各要素を除いた提案手法の有用なトピックの割合を調査した結果を**表 13** に示す．表 13 の提案手法における有用なトピックの割合の平均値が，各要素を除外した場合と比較して向上していることから，各要素が提案手法の精度に寄与していることが分かった．特に，投稿人数変化スコアを除外した場合の有用なトピックの割合が低いことから，ライフイベントの前後および特定の時期における投稿人数の増加をとらえることは，有用な情報を含むブログ記事を提示する上で，重要な要素であると明らかになった．また，「子供の小学校入学」イベントでのみ，ノイズフィルタリングを除外した場合の有用なトピックの割合が，提案手法を上回った．これは，提案手法では，「子供の小学校入学」イベントにおいて，実験参加者によって有用と判断された「子供の病気」トピックが，事前にフィルタリングされたためである．

5.5.2　投稿人数変化スコアで抽出できた有用なトピック

本研究では，3.3.1 項で述べたとおり，投稿人数変化スコアの計算に，ライフイベントの前後における投稿人数の平均の差と，ライフイベント後の月における最大の投稿人数とその他の月の投稿人数の平均との差を利用している．トピックの投稿人数の変化に関するこれらの 2 つの特徴を考慮することで，ユーザの生活の変化を反映したトピック

表 13　提案手法から各要素を除外した手法による有用なトピックの割合
Table 13　Rates of useful topics by removing each module in proposed method.

ライフイベント	提案手法	投稿人数変化スコアを除外	共起スコアを除外	ノイズフィルタリングを除外
「出産」	**0.80**	0.20	**0.80**	0.60
「就職」	**0.60**	0.40	0.20	0.40
「結婚」	**0.40**	0.00	0.20	0.20
「大学入学」	**0.80**	0.60	0.60	0.60
「子供の小学校入学」	0.60	0.00	0.40	**0.80**
平均	**0.64**	0.24	0.44	0.52

と，ユーザの多くが特定の時期に体験する興味や行動を反映したトピックの両方を抽出できたことを，以下で示す．

ライフイベントの前後における投稿人数の平均の差を利用することで，「出産」イベントにおける「育児」トピック，「子供の小学校入学」イベントにおける「小学校生活」トピックなどの，ユーザの生活の変化を反映したトピックを抽出することができた．これらのようなトピックは，ライフイベントの以前だけ，あるいは以後だけに着目すると，投稿人数の変化の幅は大きくないが，ライフイベント以後にわずかに投稿人数が増加するため，バーストに着目した比較手法では抽出することができない．なお，「結婚」イベントにおける「結婚式の準備」トピックも，投稿人数の平均の差で抽出できている．これは，結婚式は多くのユーザが体験するが，タイミングがユーザによって異なることから，投稿人数がわずかに増加し一定の人数となるためである．

また，ライフイベント後の月における最大の投稿人数とその他の月の投稿人数の平均との差を利用することで，「出産」イベントにおける「陣痛」トピックや，「大学入学」イベントにおける「テスト・レポート」トピックなどのユーザの多くが特定の時期に体験する興味や行動を反映したトピックを抽出することができた．これらのトピックは，特定の時期に投稿人数が急増するトピックであり，その特徴から，バーストが検出されやすいため，比較手法でも一部は抽出することができている．

5.5.3　共起スコアで抽出できた有用なトピック

本研究では，3.3.2 項で述べたとおり，投稿人数変化スコアだけでなく，共起スコアを考慮することで，そのライフイベントに関わりの深いトピックを抽出する．共起スコアを考慮することで，抽出できた有用なトピックについて，以下で示す．

共起スコアを利用することで，「就職」イベントにおける「残業」トピックや，「子供の小学校入学」イベントにおける「子供の教材」トピックなどの，ライフイベントと関わりが深いトピックを抽出することができた．「残業」トピックは「会社生活」トピックなどと，「子供の教材」トピックは「子供の習い事」トピックなどと共起しやすいため，共起スコアが高く，提案手法で抽出することができた．これらのようなトピックは投稿人数変化スコアだけでは，抽出することができない．また，共起スコアを考慮することで，「子供の小学校入学」イベントの「子供の夏休み」トピックなどの季節性のトピックや，「大学入学」イベントの「アニメ・漫画」トピックなどのライフイベントの影響で間接的に投稿する人数が増加したと思われるトピックのスコアを下げることができた．

5.5.4　提案手法では抽出できなかった有用なトピック

提案手法では，「出産」イベントで，比較手法で抽出できた「入院準備」トピックと，「習慣」トピックの 2 つの有用なトピックを抽出することができなかった．「入院準備」トピックは，ライフイベントを経験する月に最大の投稿人数となるため，投稿人数変化スコアは高いが，入院に必要なものを羅列しているブログ記事が多く，他のトピックと共起しにくいため，提案手法で抽出することができなかった[*20]．また，「出産」イベントで抽出した上位のトピックは，すべて投稿人数変化スコアと共起スコアの両方が高く，共起スコアを考慮することの効果が見られなかった．一方，「今日，起きる，寝る」などの単語を含む「習慣」トピックは共起スコアが有意に高く，提案手法ではノイズとして事前にフィルタリングされた[*21]．他のライフイベントでは「習慣」トピックが高い割合で出現するブログ記事は，有用でない場合が多いが，「出産」イベントでは，乳児の寝かしつけ方などの情報が含まれたため，「習慣」トピックが有用と判断されたと考えられる．

以上のように，ライフイベントによっては，共起スコアが機能しないことや，フィルタリングによって有用なトピックが事前に除外されてしまう場合があることが分かった．提案手法を用いて，「出産」イベントにおける「入院準備」トピックや，「習慣」トピックを抽出するためには，ライフイベントごとに，2 つのスコアに異なる重みを設定することや，ユーザによるトピックの適合性判定などを考慮することで，解決できる可能性があるが，これらは今後の課題とする．

[*20] ただし，投稿人数変化スコアのみを利用すると，上位 10 件に含まれる．
[*21] ただし，フィルタリングを用いない場合は上位 5 件に含まれる．

6. おわりに

本研究では，ライフイベントを経験したユーザのブログ記事を分析することにより，ユーザの変化した興味や行動に関する有用な情報を含むブログ記事を提示するためのトピックを抽出する手法を提案した．具体的には，まず，時系列トピックモデルを用いて，ライフイベントを経験したユーザ集合のブログ記事に出現するトピック分布を推定した．次に，ブログ記事内において，ほとんどのトピックと共起するトピックをノイズとしてフィルタリングした．さらに，各トピックについて，ライフイベントの前後および特定の時期における投稿人数の増加を表す投稿人数変化スコアと，ブログ記事内における他のトピックとの共起した記事数を表す共起スコアを計算し，上位のトピックを抽出した．最後に，得られたトピックが高い割合で出現するブログ記事を提示した．「出産」，「就職」，「結婚」，「大学入学」，「子供の小学校入学」について，各ライフイベントを経験したことのある 6 名の実験参加者により，それぞれのトピックで得られたブログ記事に含まれる情報の有用性を評価した結果，トピックに関連するブログ記事数のバーストを考慮した比較手法よりも有効な手法であることを確認できた．

今後の課題としては，ライフイベントごとの投稿人数変化スコアと共起スコアの適切な重みの推定手法の検討と，相互情報量などを用いた，特定のライフイベントに偏って出現するトピックを考慮する手法の検討があげられる．

謝辞　本研究にあたり，多くのご助言をいただいたシンガポール国立大学の杉山一成先生と，評価実験にご協力いただいた実験参加者の方々に深く感謝いたします．本研究の一部は，科学研究費補助金基盤研究 B（課題番号 16H02913）の助成を受けて遂行された．

参考文献

[1] Barclay, L., Everitt, L., Rogan, F., Schmied, V. and Wyllie, A.: Becoming a Mother – An Analysis of Women's Experience of Early Motherhood, *Journal of Advanced Nursing*, Vol.25, pp.719–728 (1997).

[2] Blei, D.M. and Lafferty, J.D.: Dynamic Topic Models, *Proc. 23rd Int'l Conf. Machine Learning* (*ICML 2006*), Pittsburgh, PA, USA, pp.113–120 (2006).

[3] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).

[4] Burke, M. and Kraut, R.: Using Facebook after Losing a Job: Differential Benefits of Strong and Weak Ties, *Proc. 2013 Conf. Computer Supported Cooperative Work and Social Computing* (*CSCW 2013*), San Antonio, TX, USA, pp.1419–1430 (2013).

[5] Choudhury, M.D. and Massimi, M.: "She said yes!" Liminality and Engagement Announcements on Twitter, *Proc. iConference 2015*, Newport Beach, CA, USA, pp.1–13 (2015).

[6] Gibson, L. and Hanson, V.L.: 'Digital Motherhood': How Does Technology Support New Mothers?, *Proc. SIGCHI Conf. Human Factors in Computing Systems* (*CHI 2013*), Paris, France, pp.313–322 (2013).

[7] Hofmann, T.: Probabilistic Latent Semantic Indexing, *Proc. 22nd Int'l ACM SIGIR Conf. Research and Development in Information Retrieval* (*SIGIR 1999*), Berkeley, CA, USA, pp.50–57 (1999).

[8] Kim, H.D., Castellanos, M., Hsu, M., Zhai, C., Rietz, T. and Diermeier, D.: Mining Causal Topics in Text Data: Iterative Topic Modeling with Time Series Feedback, *Proc. 22nd ACM Int'l Conf. Information and Knowledge Management* (*CIKM 2013*), San Francisco, CA, USA, pp.885–890 (2013).

[9] Kleinberg, J.: Bursty and Hierarchical Structure in Streams, *Proc. 8th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining* (*KDD 2002*), Edmonton, Canada, pp.91–101 (2002).

[10] Koike, D., Takahashi, Y., Utsuro, T., Yoshioka, M. and Kando, N.: Time Series Topic Modeling and Bursty Topic Detection of Correlated News and Twitter, *Proc. 6th Int'l Joint Conf. Natural Language Processing* (*IJCNLP 2013*), Nagoya, Japan, pp.917–921 (2013).

[11] Kumar, R., Novak, J., Raghavan, P. and Tomkins, A.: On the bursty evolution of blogspace, *Proc. 12th Int'l Conf. World Wide Web* (*WWW 2003*), Budapest, Hungary, pp.568–576 (2003).

[12] McDaniel, B., Coyne, S. and Holmes, E.: New Mothers and Media Use: Associations Between Blogging, Social Networking, and Maternal Well-Being, *Maternal and Child Health Journal*, Vol.16, pp.1509–1517 (2011).

[13] Morris, M.R.: Social Networking Site Use by Mothers of Young Children, *Proc. 17th ACM Conf. Computer Supported Cooperative Work and Social Computing* (*CSCW 2014*), Baltimore, MD, USA, pp.1272–1282 (2014).

[14] Ramakrishnan, R. and Kaur, A.: Technique for Detecting Early-Warning Signals of Performance Deterioration in Large Scale Software Systems, *Proc. 8th ACM/SPEC on Int'l Conf. Performance Engineering* (*ICPE 2017*), L'Aquila, Italy, pp.213–222 (2017).

[15] Takeda, N., Seki, Y., Morishita, M. and Inagaki, Y.: Evolution of Information Needs Based on Life Event Experiences with Topic Transition, *Proc. 40th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval* (*SIGIR 2017*), Tokyo, Japan, pp.1009–1012 (2017).

[16] Wang, C., Zhang, M., Ru, L. and Ma, S.: Automatic Online News Topic Ranking Using Media Focus and User Attention Based on Aging Theory, *Proc. 17th ACM Int'l Conf. Information and Knowledge Management* (*CIKM 2008*), Napa Valley, CA, USA, pp.1033–1042 (2008).

[17] Wang, X. and McCallum, A.: Topics over Time: A non-Markov Continuous-time Model of Topical Trends, *Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining* (*KDD 2006*), Philadelphia, PA, USA, pp.424–433 (2006).

[18] Zhang, H., Kim, G. and Xing, E.P.: Dynamic Topic Modeling for Monitoring Market Competition from Online Text and Image Data, *Proc. 21th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining* (*KDD 2015*), Sydney, Australia, pp.1425–1434 (2015).

[19] Zhou, D., Ji, X., Zha, H. and Giles, C.L.: Topic Evolution and Social Interactions: How Authors Effect Re-

search, *Proc. 15th ACM Int'l Conf. Information and Knowledge Management* (*CIKM 2006*), Arlington, VA, USA, pp.248–257 (2006).
[20] 桂井麻里衣，小野峻佑：語の共起のバースト検出に基づく研究トレンドの可視化，第 9 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2017），G7-2 (2017).
[21] 関　洋平，稲垣陽一：日常的な体験を記述したブログ文書におけるライフイベントの判定，電子情報通信学会第 12 回 Web インテリジェンスとインタラクション研究会 WI2-2008-20 (2008).
[22] 田中成典，中村健二，山本雄平，柳田尚明：情報の注目度とその重要性に基づくトピックの評価指標に関する研究，情報処理学会論文誌データベース（TOD），Vol.6, No.4, pp.69–84 (2013).
[23] 水田昌孝，熊野雅仁，小野景子，木村昌弘：文書ストリームからのバースト潜在トピック抽出における t-LDA 法の性能検証，情報処理学会研究報告数理モデル化と問題解決（MPS），Vol.2010, No.10, pp.1–6 (2010).
[24] 水沼友宏，池内　淳，山本修平，山口裕太郎，佐藤哲司，島田　諭：Twitter におけるバーストの生起要因と類型化に関する分析，情報社会学会誌，Vol.7, No.2, pp.41–50 (2013).

武田 直人

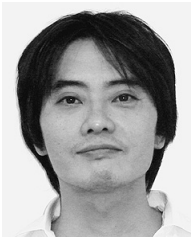
2018 年筑波大学大学院図書館情報メディア研究科博士前期課程修了．日本データベース学会会員．

関 洋平　（正会員）

1996 年慶應義塾大学院理工学研究科計算機科学専攻修士課程修了．2005 年総合研究大学院大学情報学専攻博士後期課程修了．博士（情報学）．同年豊橋技術科学大学情報工学系助手．2008 年コロンビア大学客員研究員．2018 年シンガポール国立大学客員研究員．現在，筑波大学図書館情報メディア系准教授．自然言語処理，意見分析，情報アクセスの研究に従事．ACM，ACL，電子情報通信学会，言語処理学会，日本データベース学会，人工知能学会各会員．

森下 民平

1992 年（株）シーエーシー入社．2004～2006 年スタンフォード大学コンピュータサイエンス学科客員研究員．2014 年（株）きざしカンパニー入社，現在に至る．



稲垣 陽一

1990 年東京大学文学部言語学科卒業．（株）シーエーシー入社，技術研究室に配属．1996～1998 年スタンフォード大学コンピュータサイエンス学科客員研究員．きざしサーチエンジンの研究開発を経て，2007 年 1 月より（株）きざしカンパニー代表取締役専務 CTO をつとめる．

（担当編集委員　鈴木 優）