ライフイベントを経験したユーザに共通する
トピック推移の分析手法

筑波大学
図書館情報メディア研究科
2018 年 3 月
武田　直人

目 次

表 目 次

図 目 次

第1章　はじめに

1.1　目的

「出産」や「就職」などのユーザの環境や習慣が変化する出来事（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, 17]．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 [1] や，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, 13, 14]．このような背景を踏まえ，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を分析することにより，ライフイベント経験を反映したトピックを抽出する手法を提案する．これにより，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ことができる．

表 1.1 に，有用な情報を含むブログ記事の例を示す．表 1.1 の下線部の情報のように，これからその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と思われる情報を含むブログ記事を,「有用な情報を含むブログ記事」とする．たとえば，表 1.1 の「出産」に関するブログ記事には，授乳中の乳児の噛みつき行動の対策に関する情報が含まれている．このブログ記事は，ユーザが「出産」イベントの後に持つ興味や行動を反映した「育児」トピックにより提示できる．また,「結婚」に関するブログ記事には，都民共済に加入することで，一般的な結婚式よりも費用を節約できる，という情報が含まれている．このブログ記事は，ユーザが「結婚」イベントの直前やその後に持つ興味や行動を反映した「結婚式の準備」トピックにより提示できる．さらに，これらのトピックは，ライフイベントと関わりが深く，ユーザが同一の時期に他のトピックと共起させて書く場合が多い．たとえば,「育児」トピックが出現したブログ記事中には,「乳児の成長」トピックなどが同時に出現しやすく,「結婚式の準備」トピックは,「購入」トピックなどが同時に出現しやすい．本研究では，これからライフイベントを経験するユーザに

とって有用なブログ記事を提示できるトピックは,「育児」トピックや「結婚式の準備」トピックのように，ライフイベントによる興味や行動の変化を反映しており，かつ，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と仮定する．

表 1.1: 有用な情報を含むブログ記事の例

ライフイベント	ブログ記事の一部
「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 …．...

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対象とした先行研究では，ライフイベントを経験したユーザ集合のSNS上での投稿を収集し，事前に調査する興味や行動を設定した上で，興味や行動の変化を分析しているものがある（「婚約」[5],「失職」[4]）．しかし，これらの研究は，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．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，時系列トピックモデルにより，トピックを自動的に推定する．また，推定したトピックの出現傾向を表す時系列について，ライフイベントを機に変化する度合いをスコアとして考慮する．さらに，他のトピックと共起させて書かれる度合いをスコアとして考慮することで，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ており，かつ，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の抽出を行う．

提案手法では，まず，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[1]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，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．また，提案手法と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で提示したそれぞれのブログ記事について，各ライフイベントを経験したことのある実験参加者が有用性を評価する．実験参加者により，有用と判断されたブログ記事を提示できたトピックの割合を比較することで，提案手法の有効性を確認する．

[1]特定のユーザが一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

1.2 論文の構成

本論文の構成を以下に示す．2章では，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関する研究，時系列トピックモデルを用いた研究，時系列変化の特徴やトピック同士の共起に着目したトピック抽出に取り組んでいる関連研究について論じ，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に，ライフイベント経験を反映した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，今後の課題を述べる．

第2章　関連研究

本研究では，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に，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，ライフイベント経験を反映したトピックを抽出する手法を提案する．そのため，2.1 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について論じ，本研究の新規性と有用性を明らかにする．また，本研究では，時系列トピックモデルを利用して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推定する．そのため，2.2 節では，時系列トピックモデルを用いて分析を行っている研究について論じ，本研究における時系列トピックモデルの利用について述べる．さらに，本研究では，時系列トピックモデルを利用して推定したトピックのうち，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．そこで，2.3 節では，トピックの時系列変化の特徴を利用して，特定のトピックを抽出する研究について論じ，本研究で着目する時系列変化の特徴との違いを明らかにする．最後に，本研究では，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，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る．そのため，2.4 節では，ブログ記事におけるトピック同士の共起に着目した研究について論じ，本研究でどのようにトピック同士の共起を利用するかについて述べる．これらの研究を踏まえて，2.5 節で本研究の位置付けを述べる．

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，我々が提案した興味や行動の変化を自動抽出する研究について論じる．我々の提案した研究は，先行研究と比較して，ライフイベント前後におけるトピックの出現確率の差を利用することで，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事前に設定することなく，抽出できる点で有用性

がある．

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[2] 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．まず，Twitter のハッシュタグ「#engaged」を用いて，婚約を宣言しているツイートを収集し，複数人のアノテータにより，ユーザが実際に婚約を宣言していることを確認する．このようにして得られたユーザ集合のツイートを婚約宣言前と婚約宣言後に分割し，単語や投稿内容の推移を分析する．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．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推定する．

2.1.2 興味や行動の変化を自動抽出する研究

著者らは，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，トピック推移を推定し，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抽出するための手法を提案した [17, 24]．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．図 2.1 に，著者らが提案した手法の概要を示す．

著者らが提案した手法では，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のトピックの生起確率の推移を推定する．次に，ライフイベント前後のトピックの生起確率の平均の差と分散の差に着目したスコアを定義し，図中の赤線や緑線で示したトピックを抽出する．分析の結果,「出産」イベントでは「出産報告」トピッ

[2] https://twitter.com/

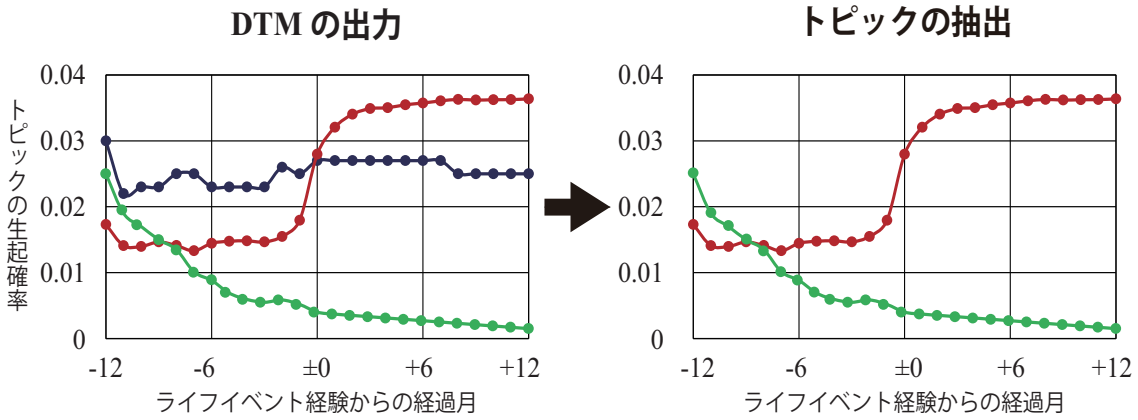

図 2.1: ライフイベントを機に生起確率が大きく変化したトピックの抽出

ク,「就職」イベントでは「就職活動」トピックなどのライフイベントの前後で生起確率が大きく変化するトピックを抽出できることを確認した.

しかし，この研究では，抽出したトピックを利用して，有用な情報を提示するための検討はされていない．また，抽出したトピックには,「就職」イベントの「就職活動」トピックなどのライフイベントの前にユーザが持つ興味や行動を反映したトピックが含まれる．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，トピックの時系列変化の特徴とトピック同士の共起に着目することで，ライフイベントの後にユーザが持つ興味や行動を反映しており，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トピックを抽出する手法を提案する．これにより，これからライフイベントを経験するユーザは，ライフイベントを経験した他のユーザの投稿した有用なブログ記事を参照できるようになる．実際に,「出産イベント」を経験するユーザは，SNS やブログにおいて，同じ境遇のユーザが投稿する記事を参照することが，多くの研究で報告されている [6, 13, 14].

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，時系列トピックモデルを用いて推定したトピックの時系列変化の分析を行っている研究について論じる.

Zhang ら [20] は，Twitter 上での商品ブランドの盛衰をリアルタイムで分析するため

に，商品ブランドに関するツイートや画像に適用できるようにDTMを拡張したトピックモデルを提案した．提案したモデルを用いた分析の結果，既存のモデルよりも，高精度に商品ブランドの盛衰を把握できることを示した．この研究では，各商品ブランドを反映したトピックに対する，Twitterユーザが同時期に持つ興味の変化をその商品ブランドの盛衰とみなし，分析している．また，Kimら[9]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．さらに，Huら[8]は，DTMを利用し，日本語のニュース記事と中国語のニュース記事に対して，時系列変化を分析し，二言語間のトピックの対応やニューストピックへの関心の差異を調査している．時系列トピックモデルで推定したトピックをスコアリングする手法としては，Wangら[18]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，トピック単位で分析を行っている．また，推定したトピックの時系列変化は，株価の変動や，ニュースによって影響を受けることが明らかにされている．本研究では，これらの研究を踏まえて，ユーザの投稿するトピックの時系列変化が，ライフイベントによって影響を受け，変化すると仮定する．そこで，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，トピックと，時期ごとのブログ記事に出現するトピック分布とを推定する．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，同じブログ記事内で同時に出現しやす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について論じる．

水田ら[22]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．LDAとの比較実験により，提案したt-LDA法は，文書内のバースト性の高いトピックを抽出する際に有効な手法であることを示した．また，Takahashiら[16]は，ニュース記事に対して，時系

列トピックモデルでトピック推移を推定し，Kleinberg のバースト検出手法 [10] を適用することで，バーストしたトピックを抽出する手法を提案した．さらに，Koike ら [11] は，Takahashi ら [16] の手法を拡張し，ニュース記事と Twitter の投稿が混合した文書ストリームに対して，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 [10] を適用する手法を提案した．

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 [12, 23, 26]．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 トピック同士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同士の共起に着目して，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，トピック同士の共起に着目した研究について論じる．

Zhou ら [21] 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，トピック同士は共起しながら，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)[19] 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同士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，有用なブログ記事に出現するトピックを，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ており，かつ，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．

2.5 本研究の位置付け

2.1.2 節で述べた手法と同様に，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，時系列トピックモデルを用いて，トピックを推定する．また，2.2 節で述べたように，時系列トピックモデルを用いて，様々なデータにおけるトピックの時系列変化を分析している研究は多数存在しているが，ライフイベントを経験したユーザのブログ記事に適用し，有用な情報を提示するための研究は，検討されていない．さらに，2.3 節と 2.4 節で述べた関連研究を踏まえ，本研究では，トピックの時系列変化の特徴と，他のトピックとの共起した記事数とを考慮したトピックのスコアリング手法を提案する．

第3章　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に，ライフイベント経験を反映したトピックを抽出する手法を提案する．図 3.1 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手法について説明する．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．

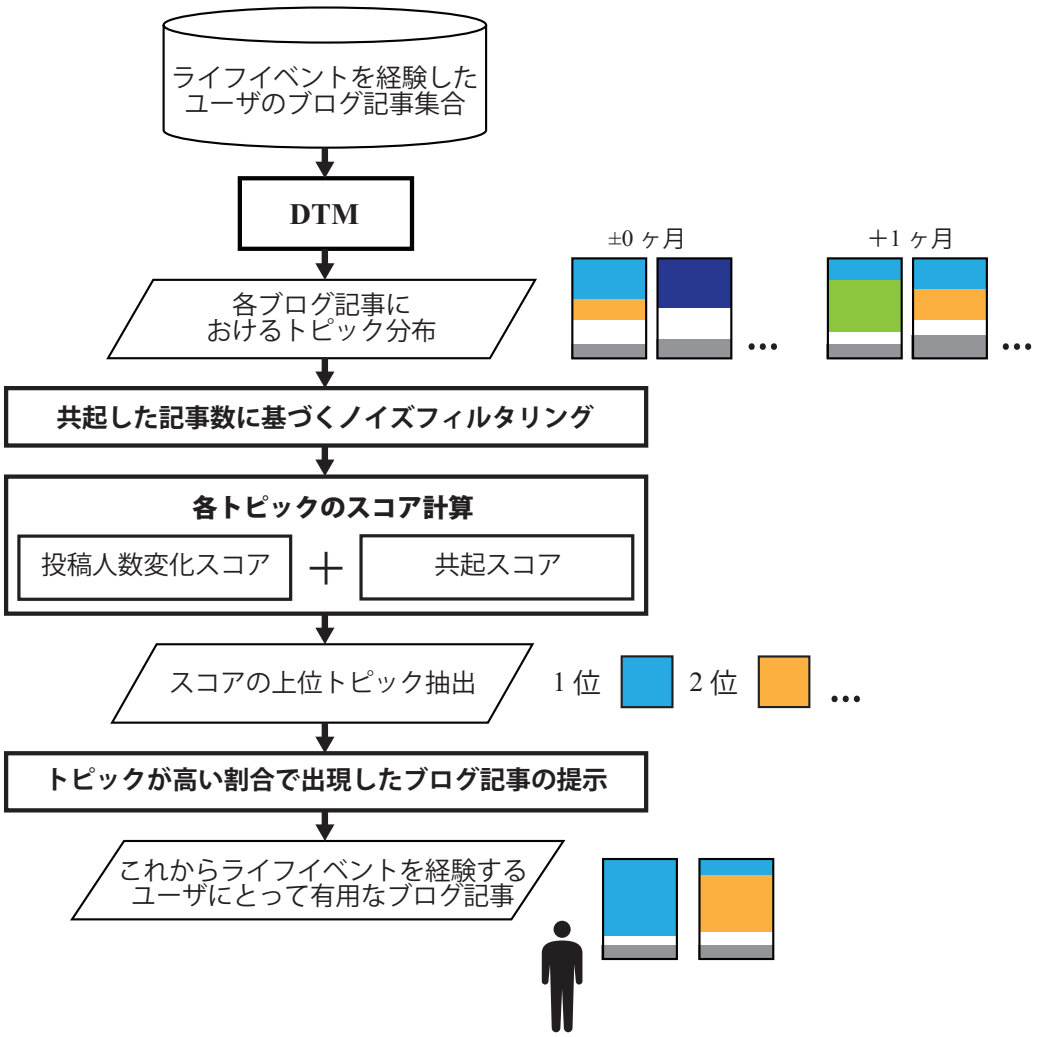

図 3.1: 有用な情報を含むブログ記事に出現するトピックの抽出

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

提案手法では，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，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，ライフイベントを経験した月を ± 0ヶ月 とし，経過月ごとに集約する．なお，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ている．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，LDA に時系列情報を加えた拡張モデルである 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．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 3.2 に，時間分割数パラメータ $TS = 3$ の DTM のグラフィカルモデルを示す．ここで，K はトピック数，D は各時刻における文書数，N は各時刻における単語数，z は各時刻におけるトピック，θ は，各時刻の文書におけるトピック分布，β は，各時刻におけるトピックの単語分布，α はハイパーパラメータをそれぞれ示す．図に示すように，DTM では，時間分割数パラメータ TS により，各トピックの経過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ヶ月間，ライフイベントを経験した月，ライフイベント以後の 12ヶ月間の計 25ヶ月間を分析対象とするため，$TS = 25$ とした．

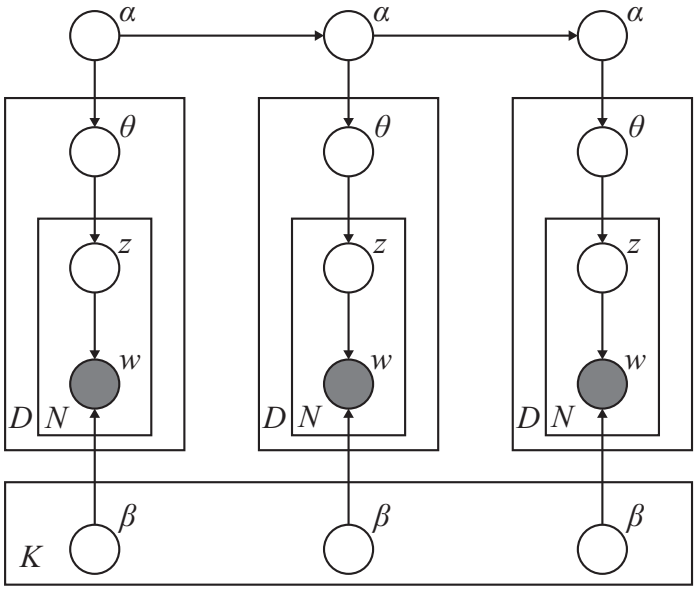

図 3.2: DTM のグラフィカルモデル（時間分割数 $TS = 3$）

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，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」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．表 3.1 に,「就職」イベントにおける「感情」トピックと「習慣」トピックに出現する代表的な単語の例を示す．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．表 3.2 に,「就職」イベントにおける「感情」トピックと「習慣」トピックが高い割合で出現するブログ記事の例を示す．

表 3.1: 「感情」トピックと「習慣」トピック中の代表的な単語

ライフイベント	トピック名	単語
「就職」	「感情」	自分，思う，言わ，考え，感じ，気持ち，思い，くれる，言葉，友達
	「習慣」	今日，時間，明日，昨日，帰っ，起き，行っ，早く，思っ，寝る

表 3.2: 「感情」トピックのブログ記事例

ライフイベント	トピック名	ブログ記事
「就職」	「感情」	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．
	「習慣」	眠たい! 今日は疲れたなー．気がついたら労働時間の 8 時間経ってたし．とりあえず，寝るー

このようなブログ記事は，投稿したユーザの個人的な体験に関するものが多く，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は含まれていないと考える．そのため，本研究では,「感情」トピックや「習慣」トピックをノイズとみなす．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同士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．

$$n(t_i) = \sum_{j:j\neq i}^{K} |d(t_i) \cap d(t_j)| \tag{3.1}$$

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，$\sum_{j:j\neq i}^{K} |d(t_i) \cap d(t_j)|$ は，トピック t_i と，K 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 [15, 26]．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．

$$n(t_i) > \mu_n + 3\sigma_n \tag{3.2}$$

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

本研究では，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なブログ記事を提示できるトピックは，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ており，かつ，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と仮定している．

そのため，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．投稿人数変化スコアを考慮することで，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ことができる．また，共起スコアを考慮することで，そのライフイベントと関わりの深いトピックを抽出することができる．

トピック t_i の投稿人数変化スコア $T\text{-}score_{t_i}$ と共起スコア $C\text{-}score_{t_i}$ 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．

$$score_{t_i} = T\text{-}score_{t_i} + C\text{-}score_{t_i} \tag{3.3}$$

以下，3.3.1 節で投稿人数変化スコア $T\text{-}score_{t_i}$，3.3.2 節で共起スコア $C\text{-}score_{t_i}$ の計算方法を説明する．

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 [17] 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 [11, 16] 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．

3.3.1.1 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ヶ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．

$$u(t_i, m)_{normalized} = \frac{u(t_i, m)}{u(m)} \tag{3.4}$$

ここで，$u(m)$ は，その月の全投稿人数を示す．

3.3.1.2 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 ヶ月 から ± 0 ヶ月 の 13ヶ月間の投稿をライフイベント前の投稿，± 0 ヶ月 から +12 ヶ月 までの 13ヶ月間の投稿をライフイベント後の投稿とみなす．

図 3.3 に，投稿人数変化スコアで捉える特徴を示す．図 3.3 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 3.3 の青線のグラフは，ライフイベント直後の ± 0 ヶ月 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めに，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する．なお，ライフイベント後の月に着目する

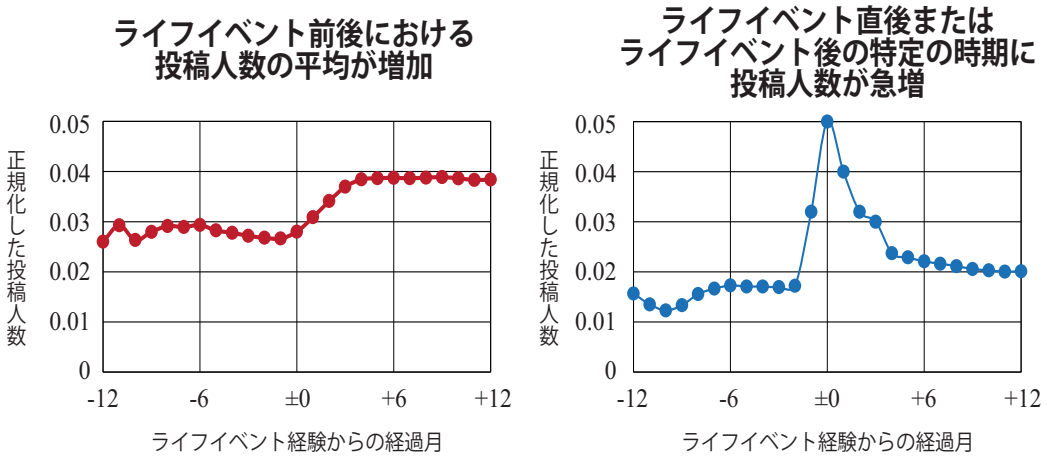

図 3.3: 投稿人数変化スコアで捉える特徴

理由は，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反映したトピックにより提示できる情報が，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と考えたためである．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．

$$m\text{-}score_{t_i} = \frac{(E(\boldsymbol{y}_{t_i}) - E(\boldsymbol{x}_{t_i})) - \mu_m}{\sigma_m} \tag{3.5}$$

$$\mu_m = \frac{1}{K}\sum_{j}^{K}(E(\boldsymbol{y}_{t_j}) - E(\boldsymbol{x}_{t_j})) \tag{3.6}$$

$$\sigma_m = \sqrt{\frac{1}{K}\sum_{j}^{K}((E(\boldsymbol{y}_{t_j}) - E(\boldsymbol{x}_{t_j})) - \mu_m)^2} \tag{3.7}$$

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 (3.4) 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．

$$r\text{-}score_{t_i} = \frac{(\max(\boldsymbol{y}_{t_i}) - E(\boldsymbol{u}_{t_i} \setminus \arg\max(\boldsymbol{y}_{t_i})) - \mu_r}{\sigma_r} \tag{3.8}$$

$$\mu_r = \frac{1}{K}\sum_{j}^{K}(\max(\boldsymbol{y}_{t_j}) - E(\boldsymbol{u}_{t_j} \setminus \arg\max(\boldsymbol{y}_{t_j})) \tag{3.9}$$

$$\sigma_r = \sqrt{\frac{1}{K}\sum_{j}^{K}(\max(\boldsymbol{y}_{t_j}) - E(\boldsymbol{u}_{t_j} \setminus \arg\max(\boldsymbol{y}_{t_j}) - \mu_r)^2} \tag{3.10}$$

ここで，トピック t_i のすべての月に対する，式 (3.4) 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他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．

$$T\text{-}score_{t_i} = \max(m\text{-}score_{t_i}, r\text{-}score_{t_i}) \tag{3.11}$$

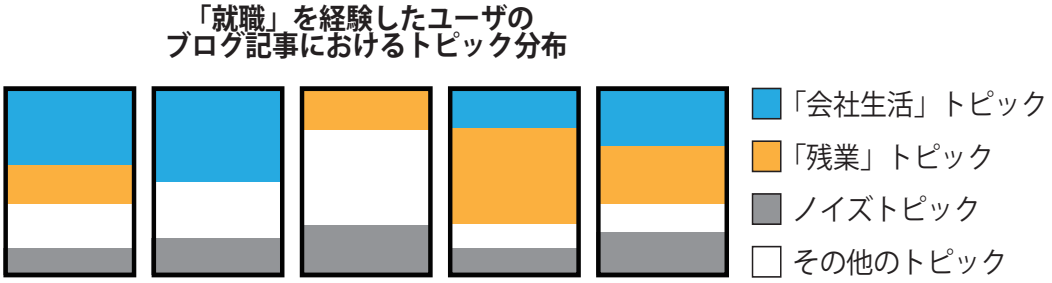

図 3.4: トピック同士のブログ記事内の共起の概要

3.3.2 トピックの共起スコア

本節では，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．提案手法における，トピック同士のブログ記事内の共起の概要を図 3.4 に示す．図 3.4 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[3]．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同士は，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，トピック同士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．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，ブログ記事内で単独で現れることが多く，共起スコアが比較的低い値となる．

なお，共起スコアを計算する上で,「感情」トピックや「習慣」トピックのような常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 $C\text{-}score_{t_i}$ を計算するための式を以下に示す．

[3]他に,「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．

$$C\text{-}score_{t_i} = \frac{\sum_{j:j\neq i}^{L} |d(t_i) \cap d(t_j)| - \mu_c}{\sigma_c} \tag{3.12}$$

$$\mu_c = \frac{1}{L} \sum_{j}^{L} \sum_{k:k\neq j}^{L} |d(t_j) \cap d(t_k)| \tag{3.13}$$

$$\sigma_c = \sqrt{\frac{1}{L} \sum_{j}^{L} (\sum_{k:k\neq j}^{L} |d(t_j) \cap d(t_k)| - \mu_c)^2} \tag{3.14}$$

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．

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の出現確率が高いブログ記事に，有用な情報が含まれていると仮定する．そのため，3.1 節で推定した各ブログ記事に対するトピック分布を用いて，抽出したトピックの出現確率が高い順に提示する．しかし，ブログ記事が短文の場合，抽出したトピックの出現確率が高くても，有用な情報は含まれにくい．表 3.3 に,「出産」における「育児」トピックの出現した短文のブログ記事の例を示す．

表 3.3: 「育児」トピックの短文のブログ記事例

ライフイベント	ブログ記事
「出産」	早いもので5ヶ月になりました．まだ寝返りしません． 離乳食はもう少したってからにしようかな...

このようなブログ記事は，これからライフイベントを経験するユーザにとって自明であることが多く，有用な情報は含まれていないと考える．そこで本研究では，文字数が500 文字以上のブログ記事について，トピックの出現確率が高い順に提示する．

第4章　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

本節では，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」[4],「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[5]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[17]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4.1に示す．

表 4.1: ブログ記事を抽出するためのクエリ一覧

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」

[4]結婚式を挙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．
[5]http://blogram.jp/

各ライフイベントについて，これらのクエリを含む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 [27]．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 4.2 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．

表 4.2: 「出産」のラベリング結果の例

ラベル	ブログ記事の一部
1	39 週 1 日，今日の明け方に 3100g 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．

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 4.3，表 4.4，表 4.5，表 4.6，表 4.7 に示す．

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 4.1 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,「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．これは，表 4.2 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と考える．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．

表 4.3: 「出産」における投稿時期ごとのブログ記事数とユーザ数

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

表 4.4: 「就職」における投稿時期ごとのブログ記事数とユーザ数

投稿時期	記事数	ユーザ数
-12ヶ月	993	26
-11ヶ月	1,050	27
-10ヶ月	1,042	30
-9ヶ月	1,024	33
-8ヶ月	1,334	35
-7ヶ月	1,548	43
-6ヶ月	1,776	47
-5ヶ月	1,675	49
-4ヶ月	1,703	50
-3ヶ月	1,583	58
-2ヶ月	1,922	89
-1ヶ月	2,830	100

投稿時期	記事数	ユーザ数
± 0ヶ月	2,757	100
+1ヶ月	2,599	95
+2ヶ月	1,863	93
+3ヶ月	1,810	91
+4ヶ月	1,747	90
+5ヶ月	1,596	88
+6ヶ月	1,538	86
+7ヶ月	1,387	83
+8ヶ月	1,512	83
+9ヶ月	1,526	84
+10ヶ月	1,308	84
+11ヶ月	1,096	87
+12ヶ月	1,019	64

表 4.5: 「結婚」における投稿時期ごとのブログ記事数とユーザ数

投稿時期	記事数	ユーザ数
-12ヶ月	788	39
-11ヶ月	806	47
-10ヶ月	824	50
-9ヶ月	901	53
-8ヶ月	1,084	58
-7ヶ月	1,136	60
-6ヶ月	1,219	66
-5ヶ月	1,477	74
-4ヶ月	1,614	78
-3ヶ月	1,622	88
-2ヶ月	1,727	95
-1ヶ月	1,686	98

投稿時期	記事数	ユーザ数
± 0ヶ月	1,737	100
+1ヶ月	1,566	100
+2ヶ月	1,464	92
+3ヶ月	1,424	91
+4ヶ月	1,192	86
+5ヶ月	1,187	84
+6ヶ月	1,085	83
+7ヶ月	1,036	79
+8ヶ月	936	75
+9ヶ月	955	74
+10ヶ月	970	71
+11ヶ月	1,001	72
+12ヶ月	1,168	71

表 4.6: 「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

投稿時期	記事数	ユーザ数
-12ヶ月	588	33
-11ヶ月	676	33
-10ヶ月	551	34
-9ヶ月	682	40
-8ヶ月	899	42
-7ヶ月	829	44
-6ヶ月	838	46
-5ヶ月	785	50
-4ヶ月	866	55
-3ヶ月	781	61
-2ヶ月	1,380	84
-1ヶ月	1,784	98

投稿時期	記事数	ユーザ数
± 0ヶ月	2,144	100
+1ヶ月	1,951	100
+2ヶ月	1,673	100
+3ヶ月	1,544	100
+4ヶ月	1,402	96
+5ヶ月	1,356	94
+6ヶ月	1,327	94
+7ヶ月	1,131	90
+8ヶ月	1,108	85
+9ヶ月	1,064	83
+10ヶ月	991	83
+11ヶ月	998	84
+12ヶ月	847	67

表 4.7: 「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

投稿時期	記事数	ユーザ数
-12ヶ月	703	33
-11ヶ月	683	34
-10ヶ月	638	34
-9ヶ月	740	39
-8ヶ月	812	43
-7ヶ月	876	53
-6ヶ月	1,116	56
-5ヶ月	1,017	59
-4ヶ月	1,070	63
-3ヶ月	1,073	73
-2ヶ月	1,399	91
-1ヶ月	1,575	99

投稿時期	記事数	ユーザ数
± 0ヶ月	1,874	100
+1ヶ月	1,776	99
+2ヶ月	1,638	98
+3ヶ月	1,692	96
+4ヶ月	1,586	96
+5ヶ月	1,545	94
+6ヶ月	1,501	90
+7ヶ月	1,381	90
+8ヶ月	1,528	87
+9ヶ月	1,459	90
+10ヶ月	1,234	86
+11ヶ月	1,148	86
+12ヶ月	1,009	67

第5章　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　目的

提案手法により抽出したトピックが，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なトピックであるかを検証するために，4節で構築したライフイベントを経験したユーザのブログ記事集合を用いて，評価実験を行う．また，トピックに関連する記事数のバーストを考慮した手法を用いて，比較実験を行い，提案手法の有効性を示す．

以下で，実験方法，比較手法，実験環境，結果について述べ，考察を行う．

5.2　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の出現確率が高いブログ記事を，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，提案手法と比較手法について，抽出した 5 件のトピックのうち，有用なトピックの割合を評価することで，どちらの手法が，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できているかを検証する．

5.3 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 [25]．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 [10] 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捉える場合，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捉える際に，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．なお，提案手法で用いた投稿人数の時期ごとの変化を利用した場合，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．そこで，トピックの出現確率の和の変化を利用した場合 [11, 16] と比較し，より有用なトピックが抽出できていると著者が判断したため，記事数の変化を利用した．

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．

5.4 実験環境

DTM の実装には，Python のライブラリである gensim[6] 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み合わ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．

$$JSD(P \parallel Q) = \frac{1}{2}(\sum_i P(i) \log \frac{P(i)}{M(i)}) + \frac{1}{2}(\sum_i Q(i) \log \frac{Q(i)}{M(i)}) \tag{5.1}$$

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，全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．

ブログ記事の形態素解析には MeCab[7] を用いた．形態素解析の辞書は，mecab-ipadic-NEologd[8] 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．

5.5 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 5.1 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

[6] https://radimrehurek.com/gensim/
[7] http://taku910.github.io/mecab/
[8] https://github.com/neologd/mecab-ipadic-neologd

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 5.2 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，各ライフイベントにおいて，提案手法により抽出できた有用と判断された一部のトピック中の代表的な単語を，表 5.3 に示す．

さらに，有用な情報を含むと判断されたブログ記事の例を表 5.4 に示す．提案手法を用いることで，表 5.4 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．

表 5.1: 有用なトピックの割合

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．

表 5.2: 提案手法と比較手法で抽出したトピック

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」，「会社生活」，「一人暮らしの料理」,「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」，「卒園・入学」，「小学校生活」,「子供の習い事」,「食事」
	比較手法	「子供の夏休み」

表 5.3: 提案手法で抽出できた有用なトピック中の代表的な単語

ライフイベント	トピック名	単語
「出産」	「育児」	ヶ月，離乳食，成長，最近，おもちゃ，増え，抱っこ，ハイハイ，遊び，生後
	「乳児の風邪」	風邪，体調，心配，小児科，病院，鼻水，下痢，早く，回復，大丈夫
「就職」	「残業」	残業，今月，終わら，定時，作業，仕事，最近，給料，無理，忙しい
	「会社生活」	仕事，会社，先輩，研修，同期，営業，電話，担当，新人，部署
「結婚」	「結婚式の準備」	結婚式，両親，ドレス，決め，準備，会場，披露宴，写真，友人，挨拶
	「結婚相手の条件」	仕事，結婚，関係，職場，年齢，恋愛，男性，欲しい，会社，付き合い
「大学入学」	「大学受験」	受験，試験，合格，今年，大学，去年，落ち，センター，受け，レベル
	「大学の授業」	授業，先生，大学，講義，履修，コマ，出席，単位，休講，実験
「子供の小学校入学」	「子供の教材」	勉強，問題，宿題，学習，算数，年生，国語，計算，漢字，プリント
	「小学校生活」	先生，学校，給食，授業，教室，担任，連絡，手紙，電話，書い

表 5.4: 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

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．...

5.6 考察

表 5.2 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，投稿人数変化スコアと共起スコアのそれぞれのスコアを考慮することで，抽出できた有用なトピックと，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．

5.6.1 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 節で述べたとおり，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．

5.6.2 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 節で述べたとおり，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，以下に示す．

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．

5.6.3 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」イベントの「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．これらの 2 つのトピックについて，以下で考察する．

「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．ただし,「入院準備」トピックは，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．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いことから，共起スコアが低い「入院準備」トピックを抽出することは難しかった．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．ただし,「習慣」トピックは，フィルタリングを用いない場合，上位 5 件に含まれる．他のライフイベントでは「習慣」トピックの出現確率が高いブログ記事は，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わかった．提案手法を用いて,「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．

第6章　おわりに

6.1　まとめ

「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．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や，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．そこで，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，ライフイベント経験を反映したトピックを抽出する手法を提案する．これにより，これからライフイベントを経験するユーザに，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ことができる．以上の研究の背景と目的について1章で述べた．

2章では，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関する研究，時系列トピックモデルを用いた研究，時系列変化の特徴やトピック同士の共起に着目したトピック抽出に取り組んでいる関連研究について論じ，本研究の位置付けを述べた．

3章では，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について述べた．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をノイズとして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する．投稿人数変化スコアを考慮することで，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と，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を抽出することができる．また，共起スコアを考慮することで，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．最後に，このようにして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．

4章では，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について述べた．

本研究では，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」,「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，各ライフイベントを経験したユーザと，ユーザの投稿したブログ記事を収集した．また，2名のアノテータを雇用し，著者によるラベリングの妥当性を検証した．

5章では，4章で構築した5つのライフイベントに関するブログ記事集合を用いて，各ライフイベントについて，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と評価を行った．提案手法と比較手法で抽出したトピックは,「出産」,「結婚」,「子供の小学校入学」イベントについては，出産経験があり育児中の3名の主婦,「就職」,「大学入学」イベントについては，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）が評価した．具体的には，実験参加者は，それぞれのトピックが高い割合で出現したブログ記事に含まれる情報の有用性を評価し，有用と判断されたブログ記事を提示できたトピックの割合で，提案手法と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を比較した．評価の結果，提案手法と比較手法との間で，有意水準1%の有意差が確認でき，提案手法の有効性を確認できた．

6.2 今後の課題

今後の課題としては，ライフイベントを経験したユーザの自動収集があげられる．DTMを用いることで，トピックの生起確率の変化や，トピック内の単語分布の変化を推定することができるため，これらを利用して，個人のブログ記事における単語や話題の変化に着目し，ライフイベント経験の有無を判断できる可能性がある．

また，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．現在の手法では,「大学入学」イベントと「子供の小学校入学」イベントで，似た単語分布となっている「食事」トピックが抽出されているが，どちらのライフイベントでも有用でないトピックと判断されている．このようなトピックは，どのライフイベントにも存在するため，単語分布をJS-divergenceなどの指標で比較し，トピックを同定した上で，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を検討している．

さらに，提案手法では，実験参加者によって有用と判断されたトピックのうち,「出産」イベントにおける「入院準備」トピックなどの，投稿人数変化スコアは高いが，共起スコアが比較的低いトピックを抽出できない場合がある．分析の結果,「出産」イベントで抽出した上位のトピックは，他のライフイベントと比較して投稿人数変化スコアが高い

トピックが多いことがわかった．そのため，他のライフイベントと比較することにより，各ライフイベントの投稿人数変化スコアと共起スコアの適切な重みを推定する手法を検討している．

謝辞

本論文は，著者が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に在籍中の研究成果をまとめたものです．筑波大学図書館情報メディア系准教授の関洋平先生には，主指導教員として，卒業研究から 3 年間にわたり，丁寧なご指導をいただき，研究のみならず多くのことを学ばせていただきました．心から深謝申し上げます．また，筑波大学図書館情報メディア系准教授の高久雅生先生，手塚太郎先生には，副査として丁寧にご指導いただきました．ここに，感謝の意を表します．

共同研究者として，卒業研究から 3 年間にわたり，様々な面で常に研究を支えていただいた株式会社きざしカンパニーの稲垣陽一さんと森下民平さんには，大変お世話になりました．ここに，深謝申し上げます．また，インターンシップを通じて，研究者としての姿勢と研究の面白さを再確認させていただいた NTT サービスエボリューション研究所の西村拓哉さんと戸田浩之さんに，深謝申し上げます．さらに，研究方針の助言をいただいたシンガポール国立大学計算機科学科上級研究員の杉山一成先生，英語プレゼンテーションの指導をいただいたメリーランド大学教授の Douglas W. Oard 先生と国立情報学研究所教授の神門典子先生に，感謝の意を表します．

最後に，研究を進めるにあたり，様々な面でお世話になったコミュニケーション理解研究室の皆様と，突然のお願いにも関わらず快く協力してくださった実験参加者の皆様に感謝の意を表します．

参考文献

[1] Lesley Barclay, Louise Everitt, Frances Rogan, Virginia Schmied, and Aileen Wyllie. Becoming a Mother – an Analysis of Women's Experience of Early Motherhood. *Journal of Advanced Nursing*, Vol. 25, pp. 719–728, 1997.

[2] David M. Blei and John D. Lafferty. Dynamic Topic Models. In *Proc. of the 23rd Int'l Conf. on Machine Learning (ICML 2006)*, pp. 113–120, Pittsburgh, PA, USA, June 2006.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

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

[5] Munmun De Choudhury and Michael Massimi. "She said yes!" Liminality and Engagement Announcements on Twitter. In *Proc. of iConference 2015*, pp. 1–13, Newport Beach, CA, USA, March 2015.

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

[7] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proc. of the 22nd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 1999)*, pp. 50–57, Berkeley, CA, USA, August 1999.

[8] Shuo Hu, Yusuke Takahashi, Liyi Zheng, Takehito Utsuro, Masaharu Yoshioka, Noriko Kando, Tomohiro Fukuhara, Hiroshi Nakagawa, and Yoji Kiyota. Cross-

Lingual Topic Alignment in Time Series Japanese / Chinese News. In *Proc. of the 26th Pacific Asia Conf. on Language, Information and Computation (PACLIC 2012)*, pp. 498–507, Bali, Indonesia, November 2012.

[9] Hyun Duk Kim, Malu Castellanos, Meichun Hsu, ChengXiang Zhai, Thomas Rietz, and Daniel Diermeier. Mining Causal Topics in Text Data: Iterative Topic Modeling with Time Series Feedback. In *Proc. of the 22nd ACM Int'l Conf. on Information and Knowledge Management (CIKM 2013)*, pp. 885–890, San Francisco, CA, USA, October 2013.

[10] Jon Kleinberg. Bursty and Hierarchical Structure in Streams. In *Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2002)*, pp. 91–101, Edmonton, Canada, July 2002.

[11] Daichi Koike, Yusuke Takahashi, Takehito Utsuro, Masaharu Yoshioka, and Noriko Kando. Time Series Topic Modeling and Bursty Topic Detection of Correlated News and Twitter. In *Proc. of the 6th Int'l Joint Conf. on Natural Language Processing (IJCNLP 2013)*, pp. 917–921, Nagoya, Japan, October 2013.

[12] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. On the Bursty Evolution of Blogspace. In *"Proc. of the 12th Int'l Conf. on World Wide Web (WWW 2003)*, pp. 568–576, Budapest, Hungary, May 2003.

[13] Brandon McDaniel, Sarah Coyne, and Erin Holmes. 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.

[14] Meredith Ringel Morris. Social Networking Site Use by Mothers of Young Children. In *Proc. of the 17th ACM Conf. on Computer Supported Cooperative Work and Social Computing (CSCW 2014)*, pp. 1272–1282, Baltimore, MD, USA, February 2014.

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

[16] Yusuke Takahashi, Takehito Utsuro, Masaharu Yoshioka, Noriko Kando, Tomohiro Fukuhara, Hiroshi Nakagawa, and Yoji Kiyota. Applying a Burst Model to Detect Bursty Topics in a Topic Model. In *Advances in Natural Language Processing*, Vol. 7614 of *Lecture Notes in Computer Science*, pp. 239–249, Kanazawa, Japan, 2012. Springer.

[17] Naoto Takeda, Yohei Seki, Mimpei Morishita, and Yoichi Inagaki. Evolution of Information Needs Based on Life Event Experiences with Topic Transition. In *Proc. of the 40th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2017)*, pp. 1009–1012, Tokyo, Japan, August 2017.

[18] Canhui Wang, Min Zhang, Liyun Ru, and Shaoping Ma. Automatic Online News Topic Ranking Using Media Focus and User Attention Based on Aging Theory. In *Proc. of the 17th ACM Int'l Conf. on Information and Knowledge Management (CIKM 2008)*, pp. 1033–1042, Napa Valley, CA, USA, October 2008.

[19] Xuerui Wang and Andrew McCallum. Topics Over Time: A non-Markov Continuous-Time Model of Topical Trends. In *Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2006)*, pp. 424–433, Philadelphia, PA, USA, August 2006.

[20] Hao Zhang, Gunhee Kim, and Eric P. Xing. Dynamic Topic Modeling for Monitoring Market Competition from Online Text and Image Data. In *Proc. of the 21th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2015)*, pp. 1425–1434, Sydney, Australia, August 2015.

[21] Ding Zhou, Xiang Ji, Hongyuan Zha, and C. Lee Giles. Topic Evolution and Social Interactions: How Authors Effect Research. In *Proc. of the 15th ACM Int'l Conf. on Information and Knowledge Management (CIKM 2006)*, pp. 248–257, Arlington, VA, USA, November 2006.

[22] 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.

[23] 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.

[24] 武田直人, 関洋平, 森下民平, 稲垣陽一. ライフイベントに依存したトピック推移の分析手法. 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）, C5-3, 2017.

[25] 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.

[26] 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.

[27] 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.

発表論文

査読付国際会議論文

1. Naoto Takeda, Yohei Seki, Mimpei Morishita and Yoichi Inagaki. Evolution of Information Needs Based on Life Event Experiences with Topic Transition. In *Proc. of the 40th Int'l. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2017)*, pp. 1009–1012, Tokyo, Japan, August 2017. **(Acceptance rate: 30.4%) (ACM SIGIR Student Travel Grant for SIGIR 2017)**

2. Naoto Takeda and Yohei Seki. Twitter User Classification with Posting Locations. In *Digital Libraries: Knowledge, Information, and Data in an Open Access Society*, Vol. 10075 of *Lecture Notes in Computer Science*, pp. 297–310, Tsukuba, Japan, 2016. Springer. **(Acceptance rate: 25.4%)**

査読付学術雑誌論文

1. （投稿中）武田直人, 関洋平, 森下民平, 稲垣陽一. ライフイベントの経験に有用なトピックの抽出と評価. 情報処理学会論文誌データベース（TOD）, Vol. 11, No. 2, 2018.

国内会議論文

1. 武田直人, 西村拓哉, 戸田浩之, 関洋平. 観光客の散策行動を考慮したエリア単位の行動分析. 人工知能学会第 31 回全国大会, 1L1-2in2, 2017.

2. 石川将吾, 武田直人, 関洋平. マイクロブログにおける投稿パターンを考慮したコミュニケーションスキルの分析. 情報処理学会第 79 回全国大会, 6K-02, 2017. **（学生奨励賞受賞）**

3. 武田直人, 関洋平, 森下民平, 稲垣陽一. ライフイベントに依存したトピック推移の分析手法. 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）, C5-3, 2017. **（学生プレゼンテーション賞受賞）**

4. 田中匠, 武田直人, 関洋平. 外国人観光客の相談相手となる Twitter ユーザの地域別検索. Web インテリジェンスとインタラクション研究会第 4 回ステージ発表, 2016. **（採択率: 23.8%）**

5. 武田直人, 佐藤朋美, 関洋平. 性別推定を利用した親しみやすいツイートへの言い換え. 人工知能学会第 30 回全国大会, 3H3-OS-17a-3, 2016.