ライフイベントに依存したトピック推移の分析手法

武田　直人† 　関　洋平†† 　森下　民平††† 　稲垣　陽一†††

† 筑波大学大学院　図書館情報メディア研究科　〒 305–8550 茨城県つくば市春日 1–2
†† 筑波大学　図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1–2
††† 株式会社きざしカンパニー　〒 103–0015 東京都中央区日本橋箱崎町 20–14 日本橋巴ビル 6F
E-mail: †s1621623@u.tsukuba.ac.jp, ††yohei@slis.tsukuba.ac.jp, †††{mimpei,inagaki}@kizasi.jp

あらまし　「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．本研究では，ライフイベントを経験する前後にユーザがブログに投稿した記事を分析することで，ライフイベント経験の前後で変化するユーザの興味や行動を明らかにする手法を提案する．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推移を抽出する．次に，トピックが現れる確率が大きく変化したトピックを選択し，単語分布の時間による発展を分析する.「出産」,「就職」,「結婚」の 3 つのライフイベントを対象とした実験の結果,「出産」では「子供の成長」トピック,「就職」では「会社生活」トピック,「結婚」では「結婚式」トピックなどのトピック推移を抽出でき，汎用的な手法であることを確認できた．また,「出産」における「子供の成長」トピックでは，ライフイベント経験の背景にある「離乳食」についての記事や，子供が「ハイハイ」をはじめたことを報告する記事が現れる時期を抽出することができ，提案手法の有効性を確認できた．

キーワード　ライフイベント, DTM (Dynamic Topic Models), 時系列分析

1. はじめに

「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．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から，他のユーザが興味を持つような情報とその時間的な推移を抽出することで，ライフイベント経験の背景にある情報要求の変化とその時期を把握する手法を提案する．これにより，ライフイベントを経験するユーザに対して，必要な情報を適切なタイミングで提示できるようになる．

ライフイベントを経験したユーザを対象とした先行研究では,「婚約」[4],「失職」[3] などに着目し，それぞれを経験したユーザ集合のソーシャルメディア上での投稿を収集し，興味や行動の変化を分析している．しかし，これらの先行研究では，事前に調査する興味や行動を設定している．そのため，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，ライフイベントの前後の時間帯におけるトピックの生起確率を比較することにより，事前に指定することなく，ライフイベントを機に変化する興味や行動を抽出する手法を提案する．

提案手法では，時系列トピックモデルの一種である DTM (Dynamic Topic Models) [1] を，ライフイベントを経験したユーザ集合のブログ記事に適用する．さらに，ライフイベントの発生により変化する興味や行動を明らかにするために，ライフイベントの前後における生起確率を比較し，傾向が大きく変化するトピックを選択する．これにより，ユーザの興味や行動を反映したトピックの時間的な推移を明らかにする．たとえば「出産」というライフイベントでは，出産後に「子供の成長」について関心を持つユーザは多い．こうしたトピックの時間的な推移を分析することにより,「離乳食」や子供が「ハイハイ」をはじめたことを報告する記事とその時期が抽出でき，子育てを経験しているユーザにとって必要な時期に適切な情報を提供できるようになる．

以下，2. 節では，ライフイベントを機に変化したユーザの興味や行動に着目した研究と，DTM を利用した研究について紹介し，本研究の位置付けを述べる．3. 節では，ライフイベントの前後のトピックの生起確率を比較するトピック推移の分析手法について詳細を述べる．4. 節では,「出産」,「就職」,「結婚」というライフイベントに着目し，実験データを作成し，判定者間一致度により，ラベリングの妥当性について検証する．5. 節では，4. 節で得られたデータを用いて，ライフイベントを機に変化したユーザの興味や行動の反映したトピックの選択とその推移の分析について検証する．また，得られた結果をもとに，考察する．最後に，6. 節で，本研究で得られた知見をまとめ，今後の課題を述べる．

2. 関連研究

2.1 ライフイベントを機に変化したユーザの興味や行動に着目した研究

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を紹介する．Choudhury ら [4] は,「婚約」を経験した Twitter ユーザの投稿に着目し，使用する単語や投稿内容を分析した．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，単語や投稿内容を分析する．分析の結果，婚約宣言前では “boyfriend” や “girlfriend” とい

う単語の使用割合は高いが，婚約宣言後はほぼ使われなくなり，“fiancé”, “fiancée”, “husband”, “wife” の使用割合が急増することが明らかになった．さらに，婚約宣言の前後では，結婚式に関連する投稿や，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．

Burke ら [3] 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明らかにした．

これらの研究では，調査する興味や行動を事前に設定している．本研究では，ライフイベントの前後の時間帯におけるトピックの生起確率を比較することにより，事前に指定することなく，ライフイベントを機に変化するユーザの興味や行動を反映したトピックの推移を分析する．これにより，ライフイベントに関連した興味や行動の変化とその時期を明らかにする．

2.2 DTM を利用した研究

DTM は，LDA(Latent Dirichlet Allocation) [2] に時系列情報を加えた拡張モデルである．DTM を利用することで，トピックの確率分布とトピック中の単語分布の時間による発展を追跡することができる．

Zhang ら [7] は，Twitter 上での商品ブランドについてのツイートや画像に DTM を適用するモデルを提案し，商品ブランドの盛衰を分析している．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抽出できることを示した．

また，Hu ら [6] は，DTM を利用し，日本語のニュースと中国語のニュースの時系列分析を行い，二言語間のトピックの対応や関心の差異を調査している．分析の結果，同じニュースに対する報道の姿勢の違いを抽出できることを示した．

これらの研究では，DTM を用いて様々な時系列データから，トピック推移を抽出している．本研究では，トピック推移を抽出した上で，ライフイベントの前後の時間帯におけるトピックの生起確率を比較することで，ライフイベントを機に変化するトピックを選択する．さらに，トピック中の単語の時間による発展を分析することで，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明らかにする．

3. 提 案 手 法

3.1 DTM を用いたトピック推移の抽出

本研究では，ライフイベントを機に変化するトピックを選択し，トピック中の単語分布の時間による発展を分析する．そのために，まず，ライフイベントを経験したユーザ集合のブログ記事を月単位に分割し，DTM によりトピック推移を抽出する．

LDA をはじめとしたトピックモデルは，同じトピック中に現れる単語同士の共起確率が高いことを利用し，訓練データに対し，文書が各トピックから生成される確率分布と，トピックが各単語から生成される確率分布の最適化を行うことができる．得られた文書におけるトピックの確率分布は，文書の特徴を表現し，トピックにおける単語の確率分布はトピックの特徴を表現する．LDA に時系列情報を加えた拡張モデルである DTM

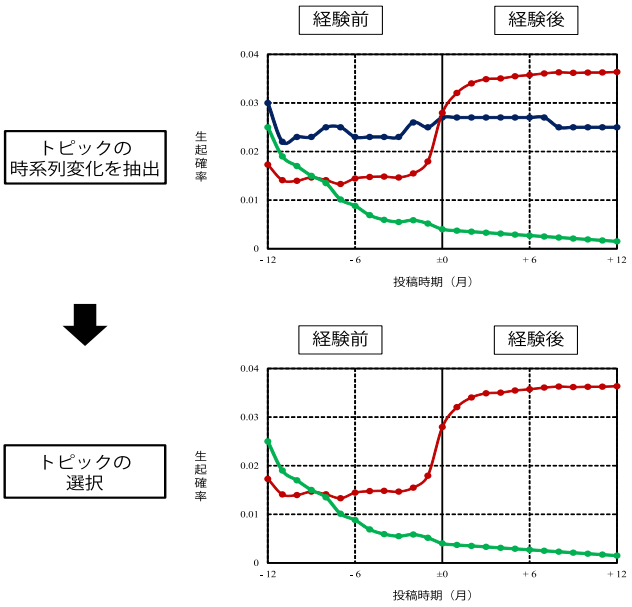

図 1　ライフイベントを機に変化するトピックの選択

では，時間分割数パラメータ TS により，トピックの確率分布は，TS 個生成される．また，各トピックについて，単語の確率分布が TS 個生成される．このうち，トピックの確率分布を追跡することで，トピックの盛衰を明らかにできる．さらに，各トピックの単語分布を追跡することで，トピック内の単語の変化を分析できる．

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，時間分割数 $TS = 25$ とした．

3.2 ライフイベントを機に変化するトピックの選択

3.1 節で述べた DTM の出力では，ライフイベント経験の前後で生起確率が変化しないトピックが含まれる．まず，ライフイベントを機に変化するトピックを選択するために，それぞれのトピックの確率分布をライフイベントを経験した月の前後に分け，各時間帯における生起確率を比較する．その際に，ライフイベントを機に生起確率が大きく変化するトピックを選択するための平均の差と，ライフイベントの前後のどちらかで生起確率が一定となるトピックを選択するための分散の差に着目した 2 つのスコアを設定し，双方のスコアの合計が高いトピックを分析対象とする．

図 1 に，ライフイベントを機に変化するトピックの選択の概要を示す．生起確率の平均の差を利用することで，ライフイベントをきっかけに，トピックの生起確率が変動したトピックを抽出できる．これは，図 1 における赤線で示したトピック推移が該当する．たとえば，「就職」における「会社生活」トピックは，ユーザにとってライフイベント後にはじめて経験することであり，トピックの生起確率の平均の差は大きくなる．一方で，生起確率の分散の差を利用することで，ゆるやかに生起確率が下降し，ライフイベントを経験してから生起確率が一定になるトピックなどを抽出できる．これは，図 1 における緑線で示したトピック推移が該当する．たとえば，「就職」における「就活」

トピックが該当する．

以上の議論を踏まえ，ライフイベントを機に変化するトピック t を判別するための $score_t$ は，ライフイベントの前後のトピックの生起確率の平均の差と分散の差との和として定義する．平均の差と分散の差の 2 つのスコアは，それぞれ z スコアを用いて正規化する．平均の差のスコアと分散の差のスコアは，それぞれ以下の式における $m\text{-}score_t$，$v\text{-}score_t$ で計算する．

$$score_t = m\text{-}score_t + v\text{-}score_t \tag{1}$$

$$m\text{-}score_t = \frac{|E(\boldsymbol{x}_t) - E(\boldsymbol{y}_t)| - \mu_m}{\sigma_m} \tag{2}$$

$$v\text{-}score_t = \frac{|V(\boldsymbol{x}_t) - V(\boldsymbol{y}_t)| - \mu_v}{\sigma_v} \tag{3}$$

$$\mu_m = \frac{1}{K}\sum_k |E(\boldsymbol{x}_t) - E(\boldsymbol{y}_t)| \tag{4}$$

$$\mu_v = \frac{1}{K}\sum_k |V(\boldsymbol{x}_t) - V(\boldsymbol{y}_t)| \tag{5}$$

$$\sigma_m = \sqrt{\frac{1}{K}\sum_k (|E(\boldsymbol{x}_t) - E(\boldsymbol{y}_t)| - \mu_m)^2} \tag{6}$$

$$\sigma_v = \sqrt{\frac{1}{K}\sum_k (|V(\boldsymbol{x}_t) - V(\boldsymbol{y}_t)| - \mu_v)^2} \tag{7}$$

ここで，ライフイベントを経験する前のトピックの確率分布は $\boldsymbol{x}_t$，経験した後の確率分布は $\boldsymbol{y}_t$，$E(\boldsymbol{x})$ は確率分布 $\boldsymbol{x}$ の平均，$V(\boldsymbol{x})$ は確率分布 $\boldsymbol{x}$ の分散，K はトピック数，μ_m と μ_v はライフベントの前後におけるトピックの生起確率の平均の差と分散の差の平均値，σ_m と σ_v はライフイベントの前後におけるトピックの生起確率の平均の差と分散の差の標準偏差をそれぞれ表す．

4. 実験データの作成

本節では，実験データの作成について述べる．今回の分析では，「出産」，「就職」，「結婚」の 3 つのライフイベントを対象として実験を行う．

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(注1) に登録されている全ブログ記事のうち，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．

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，ユーザのブログ記事から人手でライフイベント経験の有無をラベリングする．まず，ブログ記事集合から，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，ラベリングを行う．各ライフイベントで利用するクエリの一覧を表 1 に示す．各ライフイベントについて，これらのクエリのいずれか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大量のノイズが含まれる [8]．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．

表 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」,「入社式」
「結婚」	「結婚しました」,「入籍しました」

表 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 5 人目の男の子を出産しました。
0	私の飼っているキャンティーが今朝、3 匹の赤ちゃんを出産しました。

「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 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．

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 3，表 4，表 5 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 1 のクエリが現れたブログ記事を選択し，他の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

(注1)：http://blogram.jp/

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

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

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

示した 3 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．この結果から，著者によるラベリングの信頼性は高いと考え，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：ライフイベントを機に変化するトピックの選択とその推移の分析

5.1 実験方法

本節では,「出産」,「就職」,「結婚」の 3 つを対象として，ライフイベントを機に変化するトピックの選択とその推移の分析について検証する．まず，4. 節で述べた各ライフイベントを経験したユーザのブログ記事の集合から，DTM を利用して，それぞれの時期のトピックの生起確率を抽出する．次に，3.2 節で説明した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の高いトピックの上位 5 件を選択する．こうして得られた結果に対して，以下の 2 つを分析し，提案手法の有効性を検証する．

（1） トピックの時系列変化とトピック中の代表的な単語の分析

（2） トピックの単語分布の時間による発展の分析

5.2 実験環境

DTM の実装には，Python のライブラリである gensim(注2) 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．

$$JSD(P \parallel Q) = \frac{1}{2}(\sum_i P(i) \log \frac{P(i)}{M(i)}) + \frac{1}{2}(\sum_i Q(i) \log \frac{Q(i)}{M(i)}) \quad (8)$$

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 $K = 80$ 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．また，分析する単語の品詞は名詞，動詞，形容詞とした．ブログ記事の形態素解析には MeCab(注3) を用いた．形態素解析の辞書は，mecab-ipadic-NEologd(注4) を利用した．

また，一定数以上のユーザが投稿するトピックを評価対象とするために，トピックは分析期間 ($TS = 25$) において，以下の制約を満たすものとする．

（1） そのトピックに投稿したユーザ数が u 名以上

（2） 投稿したユーザの判断は，そのユーザのトピックに関連した投稿記事数が d 件以上

（3） 投稿記事のトピックの判断は，その記事のトピックの生起確率の構成比が r 以上

今回の分析では，$u = 30$, $d = 3$, $r = 0.3$ と設定した．これらの制約を満たすトピックに対して，3.2 節で提案したスコアを計算し，上位 5 件を分析した．

(注2)：https://radimrehurek.com/gensim/

(注3)：http://taku910.github.io/mecab/

(注4)：https://github.com/neologd/mecab-ipadic-neologd

表 6 「出産」における各トピック中の代表的な単語

「子供の誕生」トピック	「就寝」トピック	「陣痛」トピック	「お腹の子供の様子」トピック	「子供の成長」トピック
出産	起き	先生	体重	ヶ月
赤ちゃん	時間	陣痛	お腹	成長
幸せ	寝る	病院	検診	最近
産まれ	泣き	痛み	週間	早い
入院	抱っこ	診察	妊娠	増え

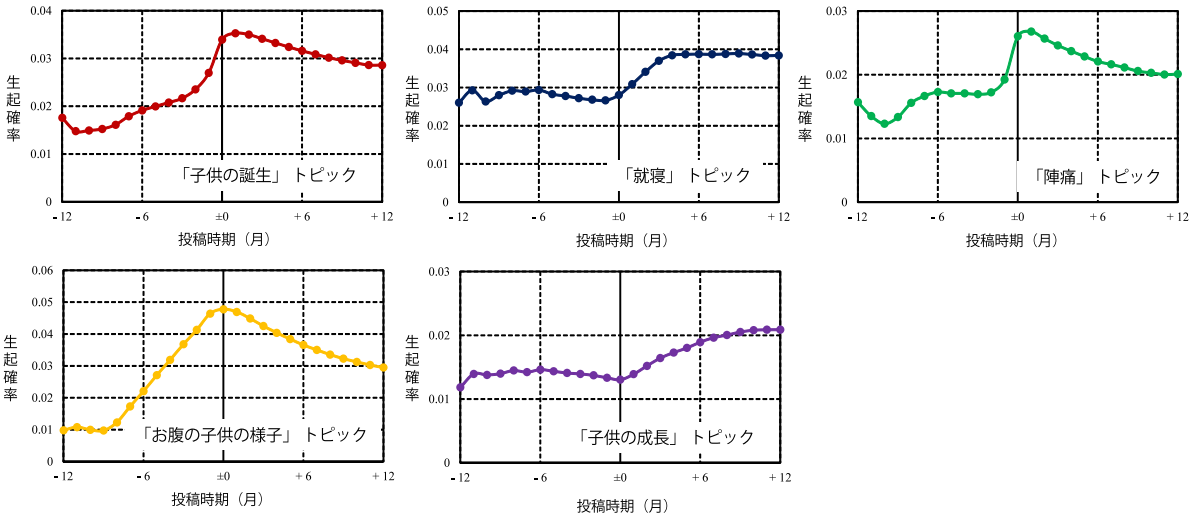

図 2 「出産」におけるスコアの上位 5 件のトピック推移

5.3 結　　果

トピックの時系列変化とトピック中の代表的な単語の分析

「出産」,「就職」,「結婚」における，スコアの高い上位 5 件のトピック中の代表的な単語を表 6，表 7，表 8 に示す．それぞれのトピックは，出現する単語から著者がトピック名を判断した．また，それぞれのライフイベントにおける，スコアの高い上位 5 件のトピック推移を図 2，図 3，図 4 に示す．図中の ± 0 は，ライフイベントが発生した月を示す．なお，図中の縦軸のスケールは，ライフイベント経験の前後の生起確率の差を表現するために，トピックごとに変更している．

図 2 と表 6 において，まず，平均の差に着目したスコアにより選択されたトピックを紹介する.「出産」の「子供の誕生」トピックは，妊娠や出産に関するトピックであり，出産を経験する月まで上昇する.「就寝」トピックは，就寝前の行動に関する記述が含まれ，出産を機に生起確率が上昇する.「陣痛」トピックは，陣痛に関するトピックであり，出産直前に急上昇し，出産後はゆるやかに下降する．このうち,「陣痛」トピックは，分散の差に着目したスコアでは，選択されないトピックである．次に，分散の差に着目したスコアにより選択されたトピックを紹介する.「お腹の子供の様子」トピックは，妊娠後の子供の様子や定期健診に関するトピックであり，出産を経験した月にピークを迎える.「子供の成長トピック」は，出産後の子供の成長に関するトピックであり，出産前は一定確率で，出産後にゆるやかに上昇する．このうち,「子供の成長」トピックは，平均の差に着目したスコアでは，選択されないトピックである．

図 3 と表 7 において，平均の差に着目したスコアにより選択されたトピックを紹介する.「就職」の「会社生活」トピックは，社会人としての生活に関するトピックであり，就職を機に急増し，一定確率となる.「感動」トピックは，上位 30 単語に「感動」,「嬉しい」などの自身の感情に関する単語が含まれており，徐々に上昇している.「購買」トピックは，様々な商品を購入した報告が多く，就職によって増加している．このうち,「会社生活」トピックと「購買」トピックは，分散の差に着目したスコアでは，選択されない．次に，分散の差に着目したスコアにより選択されたトピックを紹介する.「就活・勉強」トピックは就職活動や勉強に関するトピックであり，就職の 12ヶ月前をピークに減少し，就職後には一定確率となる.「業務」トピックは，その時期に取り組んでいる業務が含まれているトピックであり,「卒論」や「仕事」などの単語が含まれる．就職前では上昇しており，就職後には一定確率となっている．このうち,「業務」トピックは，平均の差に着目したスコアでは，選択されない．

図 4 と表 8 において，平均の差に着目したスコアにより選択されたトピックを紹介する.「結婚」の「友人との交流」トピックは，友人との交流を表しており，徐々に減少している.「料理」トピックは，料理に関するトピックであり，結婚の 3ヶ月前から上昇し，結婚後に一定確率となる.「結婚式」トピックは，結婚式に関するトピックであり，5ヶ月前から上昇し，結婚 6ヶ月後に一定確率となる (注5).「結婚への考え・悩み」トピックはブログ記事を参照すると，自身や友人の結婚を受けて感じた自身

(注5)：4.1 節で述べたように,「結婚」はユーザの「入籍しました」というクエリを含むブログ記事から判断しているため，ライフイベント経験後に「結婚式」トピックが表れやすい

表 7 「就職」における各トピック中の代表的な単語

「会社生活」トピック	「感動」トピック	「購買」トピック	「就活・勉強」トピック	「業務」トピック
仕事	今日	買っ	就活	時間
会社	すごい	買い	勉強	明日
研修	わたし	買う	面接	仕事
先輩	もらっ	購入	試験	卒論
同期	たくさん	欲しい	企業	残業

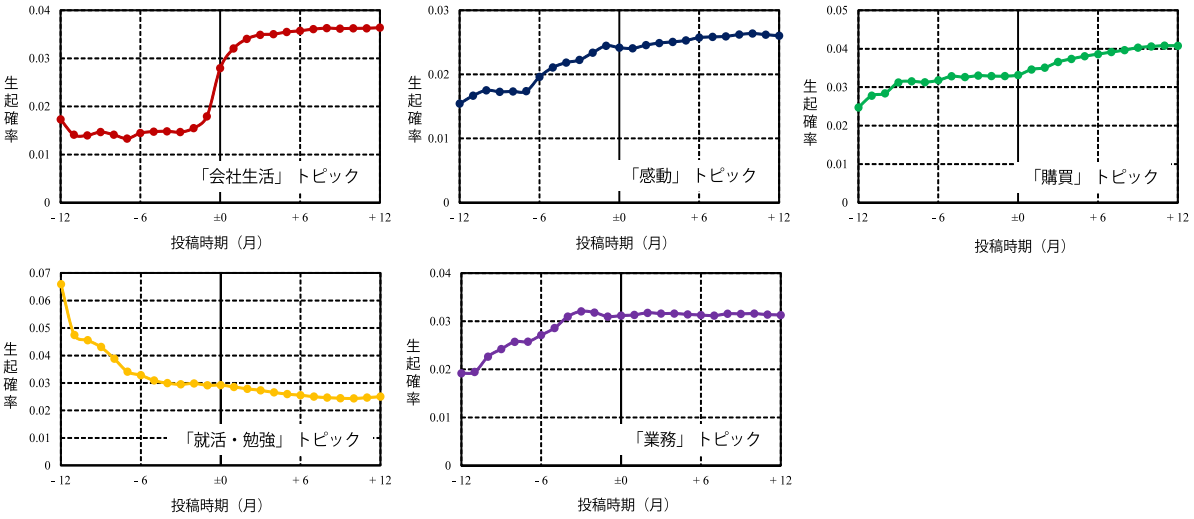

図 3 「就職」におけるスコアの上位 5 件のトピック推移

表 8 「結婚」における各トピック中の代表的な単語

「友人との交流」トピック	「料理」トピック	「結婚式」トピック	「結婚への考え・悩み」トピック	「旅行」トピック
今日	食べ	結婚式	自分	行っ
友達	ご飯	両親	思う	行き
みんな	弁当	友人	結婚	時間
昨日	野菜	結婚	仕事	旅行
行っ	今日	ドレス	女性	ホテル

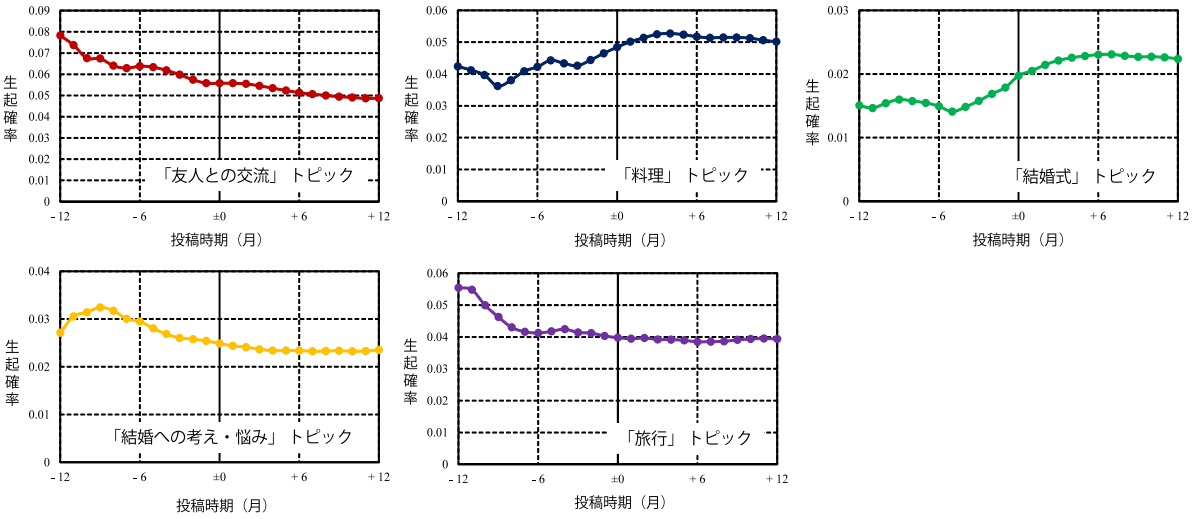

図 4 「結婚」におけるスコアの上位 5 件のトピック推移

の結婚観や恋愛観，それに伴う悩みが投稿時期に関わらず多く含まれており，徐々に減少するが，結婚後に一定確率となる．このうち,「結婚式」トピックは，分散の差に着目したスコアでは，選択されない．次に，分散の差に着目したスコアにより選択されたトピックを紹介する．また,「旅行」トピックは，旅行に関するトピックであり，12ヶ月前から減少し，6ヶ月前から一定の確率となった．

トピックの単語分布の時間による発展の分析

得られた各トピックのスコアの上位 5 トピックのうち，含まれる単語が時期によって変化しているトピックの例として,「出産」の「子供の成長」トピックと「就職」の「会社生活」トピックの単語分布の時間による発展を分析する.「子供の成長」トピックと「会社生活」トピックにおける 2ヶ月ごとの上位 15 単語の時間による発展を，表 9，表 10 に示す．

表 9 の出産後に着目すると，出産 2ヶ月後からは「離乳食」が現れ，10ヶ月後からは,「ハイハイ」が現れていることがわかる．また，表 10 では,「バイト」という単語が就職を機に現れなくなり，就職後には「上司」や「営業」という単語が現れる．

5.4 考　　察

実験の結果を踏まえて，ライフイベントの前後における生起確率の平均の差と分散の差に着目することで,「出産」では「子供の成長」トピック,「就職」では「会社生活」トピック,「結婚」では「結婚式」トピックなどのライフイベントの前後で変化するユーザの興味や行動を反映したトピックを抽出した．これらのトピックは，平均の差と分散の差のどちらかのスコアだけでは，選択できないトピックである．これにより，提案手法は，ライフイベントを経験したことによるユーザの興味や行動の変化を抽出できる汎用的な手法であることを確認できた．

また，トピック中の単語分布の時間による発展を追跡することで，ライフイベント経験の背景にある情報要求の推移が把握できる．実際に,「出産」は，その前後でユーザが多くの情報を求めるライフイベントであることが報告されている [5]．たとえば，表 9 に示すように,「出産」の「子供の成長」トピックでは，出産後の子供の成長を時期ごとに追跡できる．今回の分析の結果，ユーザの「離乳食」に関する投稿が増える時期が出産 2ヶ月後であることや，子供が「ハイハイ」をはじめたことを報告する時期が出産 10ヶ月後からという知見を得ることができた．さらに「子供の成長」トピックを含む,「離乳食」に関するブログ記事の一部を表 11 に示す．表 11 から，出産を経験したユーザの「離乳食を開始する時期」や「離乳食を早期に与えることによるアレルギー」に関する情報要求や,「離乳食として与えるもの」を抽出できていることがわかる．このように,「子供の成長」トピックを追跡することで，子育てを経験しているユーザにとって必要な時期に適切な情報を提供できるようになる．

また，表 10 に示すように,「就職」における「会社生活」トピックの単語分布の時間による発展を分析すると，就職を機に「バイト」への興味は失われ，職場における「上司」などに関する記事が増えることがわかった.「会社生活」トピックが含まれるブログ記事には，会社への要望や，上司との相性についての悩みなどが書かれていた．このような単語分布の変化を把握することで，新入社員が持つ悩みとその時期を明らかにできるため，雇用側にとって有用な知見となる．

表 11 「離乳食」に関するブログ記事の例

投稿時期	ブログ記事の一部
+5ヶ月	さてさて、離乳食をいつ始めようか悩み中です。あまり早すぎるとアレルギーになる。っと聞いた事あるしなぁ…離乳食グッズは揃っているので準備万端ですが（また記事にしますね）6/18 に保健センターで離乳食講座があるので、それを受けてからでもいいかなぁって思ってます
+5ヶ月	昨日から離乳食にチャレンジしています。 もちろん、１０倍がゆからのスタートです☆昨日の食べっぷりはというと、すべて出しちゃいました
+4ヶ月	相変わらず夜間の授乳回数が減らないのがしんどい…。離乳食とか始めたらぐっすりになるのかな?? もう 2ヶ月後くらいからは離乳食するんだなぁ…。なんだか想像できません。

また,「出産」の「体重」という単語は,「お腹の子供の様子」トピックだけでなく「ダイエット」,「カロリー」などの単語が含まれる「ダイエット」トピックにも現れた．これらのトピックにおける「体重」は，対象が子供と自身で異なっている．提案手法を適用することで，これらのトピック推移を区別して抽出でき，トピック単位で分析することの有効性を確認できた．

最後に,「出産」における「お腹の子供の様子」トピック,「出産」トピック,「陣痛」トピックなどは，ライフイベントを経験したユーザが共通してとる行動であり，ライフイベント経験の前後で生起確率が大きく変化する．このようなトピックの数は，ライフイベントによって大きく異なる．たとえば,「結婚」は以前から交際相手と同棲しているユーザと，付き合って間もなく結婚を決めたユーザではその後の行動は異なると考えられる．これに対応するために，ライフイベントを経験したユーザを経験前の投稿をもとにクラスタリングした上で，トピック推移の抽出を行うことを今後の課題とする．

6. おわりに

本研究では，時系列トピックモデルで抽出したトピック推移から，ライフイベント経験の前後で生起確率が大きく変化するトピックを平均の差と分散の差に着目し選択することで，ライフイベントを経験したユーザの興味や行動を反映したトピック推移を抽出する手法を提案した.「出産」,「就職」,「結婚」という 3 つのライフイベントを対象とした実験の結果,「出産」では「子供の成長」トピック,「就職」では「会社生活」トピック,「結婚」では「結婚式」トピックなどのライフイベントの前後で変化するユーザの興味や行動を反映したトピックを抽出できた．これにより，提案手法は，ライフイベントを経験したことによるユーザの興味や行動の変化を抽出できる汎用的な手法であることを確認できた．

また，得られたトピック推移の単語分布の時間による発展を追跡することで，ライフイベント経験の背景にあるユーザの情報要求の変化と推移を把握できる．たとえば，提案手法を適用

表 9 「子供の成長」トピックの 2ヶ月ごとの単語分布の時間による発展

-12ヶ月	-10ヶ月	-8ヶ月	-6ヶ月	-4ヶ月	-2ヶ月	± 0ヶ月	+2ヶ月	+4ヶ月	+6ヶ月	+8ヶ月	+10ヶ月	+12ヶ月
最近	最近	最近	最近	ヶ月	ヶ月	ヶ月	ヶ月	ヶ月	ヶ月	ヶ月	ヶ月	ヶ月
ヶ月	ヶ月	ヶ月	ヶ月	最近	最近	成長	成長	成長	**離乳食**	**離乳食**	成長	成長
成長	成長	成長	成長	成長	成長	最近	最近	最近	成長	成長	**離乳食**	**離乳食**
少し	少し	くれる	くれる	くれる	くれる	くれる	くれる	**離乳食**	最近	最近	最近	最近
くれる	くれる	少し	少し	早い	早い	赤ちゃん	赤ちゃん	くれる	くれる	できる	できる	できる
始め	始め	始め	始め	少し	今日	早い	今日	今日	始め	くれる	くれる	くれる
できる	できる	できる	できる	始め	始め	今日	始め	始め	今日	今日	感じ	感じ
早い	早い	早い	早い	今日	少し	始め	早い	赤ちゃん	できる	感じ	今日	今日
今日	今日	今日	今日	増え	増え	増え	感じ	感じ	感じ	始め	赤ちゃん	始め
増え	増え	増え	増え	できる	赤ちゃん	感じ	生後	生後	赤ちゃん	赤ちゃん	始め	生え
上手	上手	毎日	毎日	毎日	毎日	できる	できる	早い	早い	早い	生え	**ハイハイ**
毎日	毎日	上手	上手	上手	できる	少し	増え	できる	上手	おもちゃ	おもちゃ	おもちゃ
見せ	見せ	見せ	感じ	赤ちゃん	感じ	毎日	抱っこ	抱っこ	生後	上手	上手	上手
笑顔	笑顔	感じ	見せ	感じ	上手	抱っこ	笑っ	上手	おもちゃ	生え	早い	早い
笑っ	笑っ	笑っ	赤ちゃん	大きく	大きく	生後	**離乳食**	笑っ	抱っこ	座り	**ハイハイ**	食べ

表 10 「会社生活」トピックの 2ヶ月ごとの単語分布の時間による発展

-12ヶ月	-10ヶ月	-8ヶ月	-6ヶ月	-4ヶ月	-2ヶ月	± 0ヶ月	+2ヶ月	+4ヶ月	+6ヶ月	+8ヶ月	+10ヶ月	+12ヶ月
仕事	仕事	仕事	仕事	仕事	仕事	仕事	仕事	仕事	仕事	仕事	仕事	仕事
会社	会社	会社	会社	会社	研修	研修	同期	会社	会社	会社	会社	会社
先輩	先輩	先輩	今日	今日	今日	今日	会社	同期	先輩	先輩	先輩	先輩
バイト	今日	**バイト**	先輩	先輩	会社	会社	先輩	先輩	同期	今日	今日	今日
今日	**バイト**	研修	**バイト**	研修	先輩	同期	今日	今日	今日	同期	同期	同期
研修	研修	同期	研修	同期	同期	先輩	研修	研修	研修	**上司**	**上司**	**上司**
社員	明日	明日	同期	明日	明日	明日	明日	明日	明日	明日	明日	明日
明日	社員	社員	明日	**バイト**	社会	社会	**上司**	**上司**	**上司**	研修	研修	研修
同期	同期	電話	社会	社会	入社	入社	電話	電話	**営業**	**営業**	**営業**	**営業**
電話	電話	社会	電話	入社	**バイト**	配属	社会	**営業**	電話	担当	担当	担当
社会	社会	内定	社員	電話	電話	電話	**営業**	社員	担当	電話	電話	電話
働い	思い	思い	内定	社員	社員	社員	配属	担当	頑張っ	部署	部署	部署
思い	働い	働い	思い	思い	思い	思い	入社	言わ	社員	頑張っ	昨日	昨日
内定	内定	同じ	入社	内定	頑張り	頑張り	社員	頑張っ	新人	職場	職場	職場
同じ	同じ	頑張り	同じ	頑張り	配属	言わ	言わ	頑張り	職場	新人	頑張っ	辞め

することで,「出産」では「子供の成長」トピックを抽出できた．さらに，トピック中の単語分布の時間による発展を追跡することで,「離乳食」に関する投稿や，子供が「ハイハイ」をはじめたことを報告する投稿が現れる時期を明らかにした．

今後の課題としては，ライフイベントを経験したユーザを経験前の投稿をもとにクラスタリングした上でトピック推移を抽出することがあげられる．また，トピックの生起確率の変動を手がかりとした，ライフイベントを経験したユーザの自動抽出について検討している．

謝　　辞

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．

文　　献

[1] David M. Blei and John D. Lafferty. Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*, pp. 113–120, Pittsburgh, PA, USA, Jun 2006.

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

[3] Moira Burke and Robert Kraut. Using Facebook after Losing a Job: Differential Benefits of Strong and Weak Ties. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW 2013)*, pp. 1419–1430, San Antonio, TX, USA, Feb 2013.

[4] Munmun De Choudhury and Micheal Massimi. “She said yes!” Liminality and Engagement Announcements on Twitter. In *Proceedings of the iConference 2015*, pp. 1–13, Newport Beach, CA, USA, Mar 2015.

[5] Lorna Gibson and Vicki L. Hanson. ‘Digital Motherhood’: How Does Technology Support New Mothers? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI2013)*, pp. 313–322, Paris, France, Apr 2013.

[6] 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 *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation (PACLIC 2012)*, pp. 498–507, Bali, Indonesia, Nov 2012.

[7] Hao Zhang, Gunhee Kim, and Eric P. Xing. Dynamic Topic Modeling for Monitoring Market Competition from Online Text and Image Data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2015)*, pp. 1425–1434, Sydney, NSW, Australia, Aug 2015.

[8] 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.