社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

# ユーザのツイート履歴に含まれる単語の出現傾向を用いたライフイベント予測モデルの構築

阿部　隼[†]　白川　真澄[†,††]　原　隆浩[†]　池田　和史[†††]　帆足　啓一郎[†††]

† 大阪大学大学院 情報科学研究科　〒 565-0871 大阪府吹田市山田丘 1 − 1
†† ハッピーコンピューター株式会社　〒 530-0021 大阪府大阪市北区浮田 1-3-14 浮田ビル 4F
††† KDDI 総合研究所, 知能メディアグループ　〒 356-8502 埼玉県ふじみ野市大原 2-1-15
E-mail: †{abe.shun,hara}@ist.osaka-u.ac.jp, ††shirakawa@hapicom.jp, †††{kz-ikeda,hoashi}@kddi-research.jp

**あらまし**　本稿では，ライフイベントを経験したユーザが投稿した過去のツイート集合に含まれる単語の出現傾向を素性としてライフイベントの予測モデルを構築する手法を提案する．出産，退院，内定，妊娠，結婚の 5 種類のライフイベントを対象とした評価実験を行い，ライフイベントごとに提案手法の予測性能を評価した．
**キーワード**　Twitter，ライフイベント，イベント抽出，イベント予測

# Construction of Life Event Prediction Model using Tendency of Word Occurrence in User's Tweet History

Shun ABE[†], Masumi SHIRAKAWA[†,††], Takahiro HARA[†], Kazushi IKEDA[†††], and Keiichiro HOASHI[†††]

† Graduate School of Information Science and Technology,Osaka University　Yamadaoka 1–1, Suita-shi, Osaka, 565–0871 Japan
†† hapicom Inc.　4F Ukita Building Ukita 1–3–14, Kita-ku,Osaka-shi,Osaka, 530–0021 Japan
††† KDDI Research, Intelligent Media Laboratory　Ohara 2–1–15,Fujimino-shi,Saitama, 356–8502 Japan
E-mail: †{abe.shun,hara}@ist.osaka-u.ac.jp, ††shirakawa@hapicom.jp, †††{kz-ikeda,hoashi}@kddi-research.jp

**Abstract**　In this paper, we propose a method for constructing life event prediction models using the tendency of word occurrences in past tweets posted by users who experienced the life events. We conducted experiments for five types of life events, i.e., birth, leaving hospital, getting a job, pregnancy and marriage, and assessed the prediction performance of the proposed method for each life event.
**Key words**　Twitter，Life event，Event detection，Event prediction

## 1. 序　　論

近年，マイクロブログを代表する Twitter では，多くのユーザが身の回りの出来事をツイート (140 文字以内のテキスト) として投稿している．そのため，Twitter は実世界の出来事を抽出するための有用な情報源となっている．ツイートには，個人が経験する出産や結婚などのライフイベントに関する情報が含まれている．ツイートを解析することにより，あるユーザが近い将来にライフイベントを経験するかどうかを事前に予測することが出来れば，そのライフイベントに関連する推薦や広告など，様々なアプリケーションに利用できる．

Twitter から実世界のイベントを抽出する研究はこれまで数多く行われており，典型的な研究としては，多数のユーザが観測するイベントに焦点を当てたものが挙げられる．例えば，Sakaki ら [1] は地震の発生検知，Li ら [2] はインフルエンザの流行検知に焦点を当てている．これらの研究に加え，最近では出産や結婚など，個人が経験するライフイベントの抽出に焦点を当てた研究も行われている．例えば，文献 [3]〜[7] はユーザの社会的な地位やツイートを行う頻度，他のユーザとの交流パターンなど様々な特徴を用いてライフイベントを抽出している．これらの研究では，ライフイベントの発生後にそのライフイベントについて言及しているツイートを捉えることを目的としている．そのため，既存手法は，近い将来に出産するユーザを予測していち早くベビー用品の宣伝を行うなど，ライフイベント

発生前に行動を起こしたい場合には適用できない．

イベントの予測という観点でみると，多数のユーザが観測するイベントのうち，開催が予定されているものを対象にした研究が行われている [8], [9]．このようなイベントは，多くのユーザがイベントの開催時期を含むツイートを投稿する可能性が高いため，これらのツイートを検出し，解析することでイベントの発生を予測出来る．しかし，本研究が対象とするライフイベントは個人が経験するイベントであり，本人あるいは少数の親密な人しか関連した投稿を行わない．そのため，ライフイベントの発生時期に言及したツイートが投稿されることがほとんどない．したがって，イベントの開催時期を含むツイートを用いた手法を適用できない．このように，既存手法ではライフイベントの予測を行うことが難しく，技術的な課題があった．

そこで本研究では，すでにライフイベントを経験した Twitter ユーザのツイート履歴をもとにライフイベントの予測モデルを構築する手法を提案した．ライフイベントが近い将来起こるであろうユーザは，一般的なユーザとは異なる習慣や行動がツイートに現れると考えられる．そこで提案手法では，ライフイベントを経験したユーザの過去のツイート履歴を，ライフイベントを近い将来に経験するユーザのツイートとして利用し，ツイートの特徴を利用した分類器を学習する．具体的にはまず，既存手法によりライフイベントを経験した一部のユーザを抽出し，人手によりノイズを取り除く．その後，ライフイベントを経験したユーザとランダムに抽出したユーザの過去のツイート集合に含まれる単語の出現傾向を素性として SVM による分類器を構築する．評価実験では，ライフイベントごとに提案手法の予測性能の評価を行い，キーワードに基づくベースライン手法との比較を行った．

以下，第 2 章では提案手法であるライフイベント予測モデルの構築について記述し，第 3 章では評価実験について述べ，最後に第 4 章で本研究のまとめと今後の課題について述べる．

## 2. ライフイベント予測モデルの構築

### 2.1 概　　要

本研究では，近い将来 (数ヶ月以内) にライフイベントが発生するユーザを検出することを目的とし，ライフイベント予測モデルを構築する．提案手法では，ライフイベントを経験したユーザの過去のツイートを，ライフイベントを近い将来に経験するユーザのツイートとみなし，予測モデルの構築に利用する．ライフイベントを経験したユーザ (イベントユーザ) のツイート履歴を遡ってツイートを収集，解析し，ランダムに抽出したユーザ (一般ユーザ) と比較することで，ライフイベントの予測モデルを構築できると考えられる．提案手法の具体的な処理の流れを図 1 に示す．図 1 の各処理は以下の通りである．

- **イベントユーザ抽出 (2.2 節):**フォロワーのリプライに基づく手法 [7] により，ライフイベントを経験したユーザを抽出し，人手でノイズを除去する．
- **過去のツイート収集 (2.3 節)** :イベントユーザがライフイベントを経験した時点から 6 ヶ月遡ってツイートを取得する．また，イベントユーザとの比較として用いるため一般ユーザの 6 ヶ月分のツイートも取得する．
- **ツイート集合の特徴抽出 (2.4 節)** :イベントユーザと一般ユーザのツイート集合の性質の違いを表現するため，ツイート集合のテキストを解析し，特徴を抽出する．
- **分類器の構築 (2.5 節)** :イベントユーザのツイート集合を正例，一般ユーザのツイート集合を負例として，それぞれ抽出した特徴を利用して SVM 分類器を構築する．

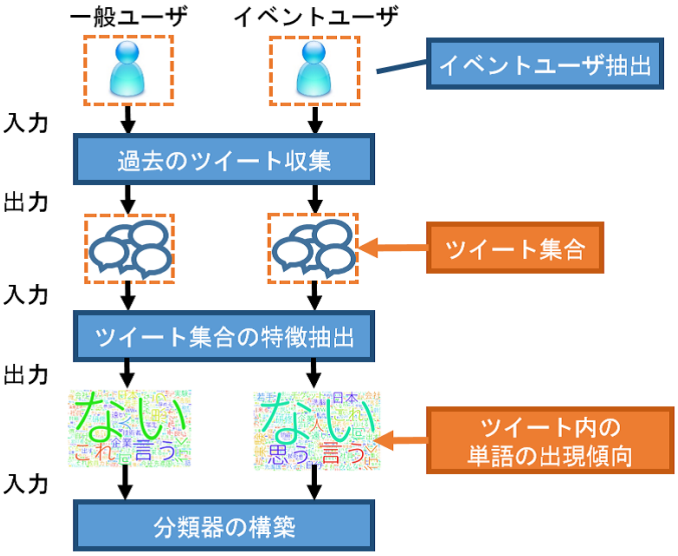


図 1　提案手法の処理の流れ

以下の節では，上記の各処理について詳しく記述する．

### 2.2 イベントユーザ抽出

提案手法では，イベントユーザがライフイベントを経験した時点からツイート履歴を遡って過去のツイートを収集，解析することでライフイベント予測モデルを構築する．そのため，まずイベントユーザを抽出する必要がある．そこで，本研究では，ライフイベント抽出手法の中でも比較的高精度な，フォロワーのリプライに基づく手法 [7] によりライフイベントを抽出する．

文献 [7] では，ユーザが自身のライフイベントについて述べている場合，そのツイートに対し，フォロワーから「おめでとう」などの定型表現を含むリプライが送られる傾向にあることを利用している．なお，ユーザ自身のライフイベントについて述べているツイート全てに「おめでとう」などを含むリプライが送られるとは限らない．一方で，投稿者のライフイベントとは関係のないツイートに対しては，フォロワーから「おめでとう」などを含むリプライは付きにくい．そのため，この手法により，実際にライフイベントを経験したユーザによるツイートを高精度で抽出でき，ライフイベント予測モデルの教師データとして利用できる．

フォロワーのリプライに基づく手法の具体的な処理の流れを説明する．本研究では，定型表現「おめでとう」を含むリプライを取得する．次に，得られたリプライに対し，参照している元のツイートを取得する．その後，抽出したいライフイベントを示すキーワードを指定し，リプライが参照している元のツ

イートの中からそのキーワードを含むツイートをフィルタリングし，出力する．このとき，出力されるツイートにはユーザ自身のライフイベントについて述べていないものも含まれるが，出力の多くはユーザ自身のライフイベントに言及したツイートであるため，十分な量の正例データを集めることは難しくない．そこで，人手によりノイズを除去し，自身のライフイベントに言及したツイートを取得する．

### 2.3 イベントユーザの過去のツイート収集

2.2 節で抽出したイベントユーザについて，ライフイベントを経験した時点 (自身のライフイベントの発生について言及したツイートが投稿された時点) からツイート履歴を遡り，過去のツイートを収集する．これにより「ライフイベントを近い将来に経験するユーザのツイート集合」を取得できる．あるユーザの現在を起点とし，将来ライフイベントが起こるかを考えても，実際にライフイベントが起こるかどうかがわからないが，ライフイベントが起こった時点から過去のツイートを収集すれば，過去を起点として考えたとき，将来ライフイベントが起こるユーザによるツイート集合と見なすことができる．

対象とするイベントユーザの過去のツイートは Twitter REST API(注1) を用いて収集する．この API は 2016 年 10 月時点において，ユーザの最新のツイートから 3200 件のみ遡って収集できる．そのため，ツイートの頻度が極端に多いユーザの場合，正例データの作成に必要な期間を遡ることができない．そこで，十分な期間 (本研究では 6 ヶ月とした) のツイートを遡れなかったユーザは除去する．また，ほとんどツイートを投稿しないユーザの場合，ツイート履歴を利用した予測モデルを構築する上で精度低下の要因となる．そこで，6 ヶ月遡った際のツイート数が一定数以下 (本研究では 60 とした) のユーザを取り除く．また負例データを作成するため，一般ユーザについても同様の処理によりツイート履歴を遡り，過去のツイートを収集する．

### 2.4 ツイート集合の特徴抽出

2.3 節の処理により，ライフイベントを近い将来に経験するユーザ，および経験しないユーザのツイート集合をそれぞれ取得できる．これらのツイート集合を用いて，ライフイベントを近い将来に経験するユーザにのみ現れる特徴を抽出する．ライフイベントを近い将来に経験するユーザは，一般ユーザと比べて，自身のライフイベントに関連したツイートを投稿する可能性が高く，ツイート内の単語の出現傾向にも特徴が現れると考えられる．そこで本研究では，単語の出現を特徴として利用する．

具体的には，ツイート集合に含まれる単語を各次元とし，単語の出現確率のベクトルとして特徴を表現する．単語の出現確率は，ツイート集合に含まれるツイート数のうち，その単語を含むツイート数の割合として表す．なお，近い将来にライフイベントを経験するユーザであっても，自身のライフイベントに関連したツイートを投稿する割合は通常少ないため，単語の出現確率をそのまま用いた場合，一般的な語の特徴が支配的になる．そこで，単語の出現確率を計算した後，ランダムに抽出したツイート群の単語の出現確率を減算する．これにより，近い将来にライフイベントを経験するユーザのツイートに現れやすい特徴的な単語を抽出できる．例えば，出産では出現確率の高い単語の上位に「陣痛」,「赤ちゃん」,「病院」,「妊娠」,「検診」などが出現しやすくなる．ランダムに抽出したツイート群から算出した出現確率を用いて調整することで，確率の高い単語の上位に特徴的な語が集まるため，確率の高い上位 300 件の単語を特徴として利用する．また一般ユーザに対しても同様に，単語の出現確率を計算した後，ランダムに抽出したツイート群の単語の出現確率を減算する．

### 2.5 分類器の構築

2.4 節で抽出した特徴をもとに，ライフイベントが近い将来起こるユーザか否かを予測する分類器を構築する．分類に用いる学習アルゴリズムとして Support Vector Machine (SVM) を用いる．SVM とはパターン識別モデルの一種で，2 つのグループに分類する問題によく使われる分類器である．以下では，SVM 分類器を用いてライフイベント予測を行うにあたって，学習時や判定時の入出力形式をどのように定義するかについて述べる．

ライフイベント予測では，将来ライフイベントを経験するであろうユーザであっても，それがどの程度先のことなのかがわかる必要があるため，指定した期間内にライフイベントが発生するかどうかを予測する必要がある．また，ライフイベント予測モデルを構築する際にはライフイベントを経験した時点が与えられているため，あるツイート集合がライフイベント発生前のいつの時点のものなのか把握できる一方，判定時には入力となるツイート集合がライフイベント発生前のいつの時点のものか，あるいはライフイベントが発生しないユーザによるものなのかが区別できない．そこで，本研究で提案するライフイベント予測モデルでは，あるユーザの固定期間 (本研究では 1 ヶ月とする) のツイート集合を入力とし，そのユーザが予測期間 $K$ ヶ月以内にライフイベントを経験するか否かを判別する．したがって，対象のユーザが $K+1$ ヶ月後にライフイベントを経験する場合は，$K$ ヶ月以内にライフイベントを経験しない，と判別される必要がある．

分類器を構築するための教師データは，ユーザのツイート履歴を固定期間ごとに区切り，それぞれの期間のツイート集合に対してラベル付けを行うことで作成出来る．本研究では固定期間を 1 ヶ月としているため，イベントユーザがライフイベントを経験した時点からツイート履歴を過去に $K$ ヶ月遡り，1 ヶ月ごとにツイート集合を区切り，正例データを作成する．また，一般ユーザから得られた適当な時点における 1 ヶ月分のツイート集合を負例データとする．

## 3. 評価実験

本章では，提案手法の予測性能の評価に関して記述する．

### 3.1 評価環境

本研究では，既存手法 [7] により抽出可能なライフイベントの中で，多くの人が生涯の中で経験すると思われる「出産」,「内

(注1)：https://dev.twitter.com/rest/reference/get/statuses/user_timeline

表 1　教師データの正例，負例のラベル数 (正例:負例)

| イベント | $K=1$ | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 出産 | 191:11460 | 382:11460 | 573:11460 | 764:11460 | 955:11460 | 1146:11460 |
| 内定 | 271:16260 | 542:16260 | 813:16260 | 1084:16260 | 1355:16260 | 1626:16260 |
| 退院 | 184:11040 | 368:11040 | 552:11040 | 736:11040 | 920:11040 | 1104:11040 |
| 妊娠 | 164:9840 | 328:9840 | 492:9840 | 656:9840 | 820:9840 | 984:9840 |
| 結婚 | 241:14460 | 482:14460 | 723:14460 | 964:14460 | 1205:14460 | 1446:14460 |

表 2　テストデータの正例，負例のラベル数 (正例:負例)

| イベント | $K=1$ | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 出産 | 48:2880 | 96:2880 | 144:2880 | 192:2880 | 240:2880 | 288:2880 |
| 内定 | 68:4080 | 136:4080 | 204:4080 | 272:4080 | 340:4080 | 408:4080 |
| 退院 | 47:2820 | 94:2820 | 141:2820 | 188:2820 | 235:2820 | 282:2820 |
| 妊娠 | 42:2520 | 84:2520 | 126:2520 | 168:2520 | 210:2520 | 252:2520 |
| 結婚 | 61:3660 | 122:3660 | 183:3660 | 244:3660 | 305:3660 | 366:3660 |

定」,「退院」,「妊娠」,「結婚」の 5 種類のイベントを評価の対象とした．データセットの作成手順として，まず 2016 年 6 月 10 日から 8 月 8 日までの「おめでとう」を含むリプライのうち，リプライが参照している元のツイートに「出産」,「内定」,「退院」,「妊娠」,「入籍」のいずれかが含まれている場合に，そのツイートの投稿日時をライフイベント発生時点としてイベントユーザを取得した．結婚のイベント発生時点は入籍した時点であると考えられるため,「結婚」ではなく「入籍」をキーワードとして用いた．次に，2.3 節で述べた手順によりイベントを経験したユーザのツイートを収集した．また，ランダムに選んだ一般ユーザに対してもイベントユーザと同様に 6 ヶ月分のツイートを取得した．ランダムに選んだユーザの中には，ライフイベントを近い将来に経験するユーザが入る可能性もあるが，割合として少ないため無視できる．

教師データ，テストデータともに，1 ヶ月分のツイート集合を 1 サンプルとして，正例または負例のラベルを付与した．教師データは，イベントユーザの過去の $K$ ヶ月までのツイート集合を正例，一般ユーザについては 6 ヶ月間のツイートを負例とした．イベントユーザの $K$ ヶ月より過去のツイート集合は教師データに利用しない．理由として，本研究では $K$ ヶ月を境界とした二値分類として問題を定式化しているが，実際には明確な境界はなく，イベントユーザの $K$ ヶ月以前のツイート集合には、正例と負例の特徴が混ざっていると考えられるためである．テストデータでは，対象のユーザが $K+1$ ヶ月後にライフイベントを経験する場合は，$K$ ヶ月以内にライフイベントを経験しない，と判別される必要があるため，イベントユーザの過去 $K$ ヶ月までのツイート集合を正例とし，一般ユーザの 6 ヶ月間のツイート集合，およびイベントユーザの過去 $K$ ヶ月以前のツイート集合を負例として扱った．表 1，2 に各ライフイベントの $K$ の値ごとの教師データ，テストデータの正例，負例の数を示す．

テストデータにおいてはスパムと思われるアカウントのツイート集合は除外した．これは，スパムアカウントは個人と紐付かないためライフイベントを予測する必要がない一方，同じ単語を含むツイートを繰り返し投稿する傾向があり，ツイート内にライフイベントに関連した単語が入っている場合に判定を誤ることが多いためである．スパムアカウントの特徴として「全てのツイートに宣伝のための URL が付けられている」,「同じツイートを繰り返し投稿する」などが挙げられる．そこで「全てのツイートに URL が付属」,「ツイート集合における 20 %以上のツイートについて，重複するツイートが存在する」,「ツイート集合中で使用されるユニークな単語数がツイート数以下」という条件を設定し，いずれかの条件に当てはまるアカウントをスパムアカウントとした．

本研究では，ライフイベントを近い将来に経験するユーザは一般ユーザとは異なる習慣や行動がツイートに現れると仮定している．ここで，その特徴がライフイベント発生前のどの期間まで顕著に現れるか，すなわち適切な予測期間 $K$ を考える必要がある．なぜなら，一般ユーザとは異なる特徴が出にくい期間まで遡って正例データを作成すると，分類器の精度低下の要因となるためである．また，ライフイベントごとに，イベント発生直前においてのみ予測可能なもの，イベント発生前の長期間において予測可能なものなど，予想しやすい期間の傾向は異なると考えられる．そこで，ライフイベントごとに予測期間 $K$ を 1 と 6 に設定し，イベント発生直前における予測 ($K=1$) と，イベント発生前の長期間における予測 ($K=6$) の精度を評価した．評価実験では，Precision-Recall 曲線を評価尺度として用い，単純なキーワードによる手法をベースラインとして提案手法との比較を行った．ベースライン手法では，1 ヶ月分のツイート集合に各ライフイベントを表す「出産」,「内定」,「退院」,「妊娠」,「入籍」の単語が含まれる数が多い順にテストデータをソートし，Precision-Recall 曲線を描くことで提案手法との比較を行った．なお，Precision-Recall 曲線では，グラフの右上に位置する (すなわち，Precision と Recall が共に 1.0 に近い) ほど性能が良いことを意味する．

Precision-Recall 曲線では異なる $K$ に対する評価結果を直接比較できないため，正例と負例のデータ数が異なる場合でも分類器の性能を比較できる ROC 曲線下の面積 (AUC) を用いた調査を行った．AUC はランダムな分類の場合に 0.5, 完全な分類の場合に 1.0 となる．ライフイベントごとに $K$ を 1 から 6 の間で変化させて $K$ の値ごとの評価結果を比較した．

### 3.2 評価結果

5 種類のイベントに対し，図 2〜6 に Precision-Recall 曲線による評価結果，表 3 に ROC 曲線下の AUC による評価結果を示す．以下では，ベースライン手法との比較評価，5 種類のライフイベント予測の評価についてそれぞれ議論する．

#### 3.2.1 ベースライン手法との比較評価

予測期間 $K$ を 6 とした際の Precision-Recall 曲線による評価結果より，出産，内定，妊娠，結婚のイベント予測においては，提案手法の曲線がほぼ常にベースライン手法の曲線より上に位置している．このことから，これらのイベントにおいてイベント発生前の長期間における予測を行う場合，提案手法がベースライン手法よりも精度良くライフイベントを予測できるといえる．ベースライン手法は，対象のユーザが固定期間内でキーワードを含むツイートを投稿しているかという情報のみに基づいて予測を行うため，該当のキーワードを使っていないケースは全て見逃す．それに対して提案手法は，イベントに関

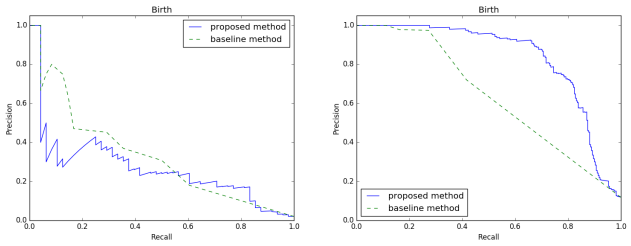


図 2　出産のイベントの予測精度の比較 (左:$K = 1$, 右:$K = 6$)

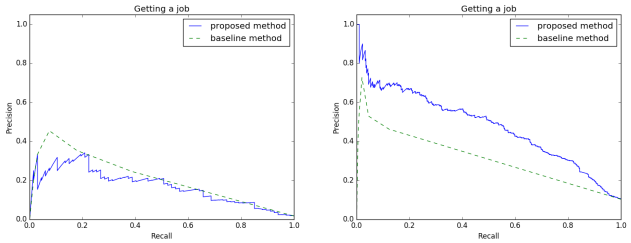


図 3　内定のイベントの予測精度の比較 (左:$K = 1$, 右:$K = 6$)

連した様々な単語を特徴として捉えられるため，将来イベントが起こるユーザをベースライン手法よりも多く捉えることが出来たと考えられる．一方で，退院のイベントの予測においては，ベースライン手法の精度が提案手法を上回っている．このことから，退院のイベントは「退院」という単語の出現回数が予測に役立っており，また，それ以外の単語はノイズとなるケースが多いと考えられる．

予測期間 $K$ を 1 とした際の Precision-Recall 曲線は，提案手法，ベースライン手法共に精度が低くなる傾向があることがわかった．これは，予測期間 $K$ を 1 とする場合，イベント発生直前 1 ヶ月以内とイベント発生の 1 ヶ月以上前で単語の出現傾向が大きく異なる場合でなければ正例と負例の分離が困難であることが原因であると考えられる．$K = 6$ の場合と同様に退院のイベントの予測のみ，ベースライン手法は提案手法より高い精度を発生している．これは，直前 1 ヶ月以内のツイートに「退院」の単語が含まれる数が 1 ヶ月以上前と比べて特に多くなる傾向にあったためであると考えられる．

以上より，提案手法はある程度長い期間の間にライフイベントを経験するか否かを予測する場合，ベースライン手法よりも高い精度を達成できる一方，1 ヶ月などの短期間の予測の場合は，単語の出現傾向を用いたライフイベント予測が難しく，ベースライン手法の性能を上回れないことがわかった．よって，次の項では，$K = 6$ としたときの各ライフイベント予測の評価について議論する．

### 3.2.2　各ライフイベントの予測の評価 ($K = 6$ の場合)

**出産のイベントの予測** (図 2 右) は Precision-Recall 曲線が他のイベントの曲線よりも上に位置しており，最も高い予測精度を達成していることがわかる．また表 3 より，AUC も $K$ の値に関わらず他のイベントと比べると最も高い．このことから，出産を経験するユーザはイベント発生前のツイートに明確な特徴が長期間現れると考えられる．出産のイベントを経験するユーザは，その前に必ず妊娠している期間 (一般的におよそ

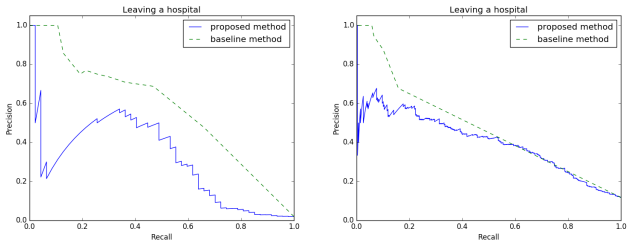


図 4　退院のイベントの予測精度の比較 (左:$K = 1$, 右:$K = 6$)

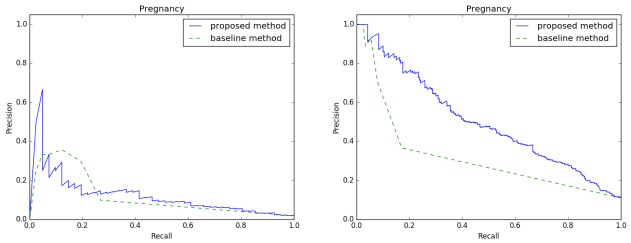


図 5　妊娠のイベントの予測精度の比較 (左:$K = 1$, 右:$K = 6$)

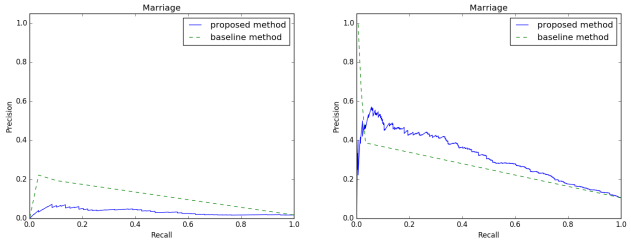


図 6　結婚のイベントの予測精度の比較 (左:$K = 1$, 右:$K = 6$)

表 3　各ライフイベントの $K$ ごとの $AUC$

| イベント | $K = 1$ | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 出産 | 0.895 | 0.878 | 0.897 | 0.913 | 0.927 | 0.925 |
| 内定 | 0.880 | 0.867 | 0.874 | 0.862 | 0.854 | 0.854 |
| 退院 | 0.837 | 0.797 | 0.797 | 0.800 | 0.802 | 0.806 |
| 妊娠 | 0.815 | 0.820 | 0.815 | 0.830 | 0.841 | 0.835 |
| 結婚 | 0.612 | 0.687 | 0.728 | 0.740 | 0.761 | 0.771 |

10 ヶ月) があるため，ほとんどのユーザに共通した特徴が長期間現れたために，最も高い予測精度を達成したと考えられる．

**内定のイベントの予測** (図 3 右) は，出産の Precision-Recall 曲線と比較すると精度が低くなっていることがわかる．原因として，内定のイベントのユーザごとのばらつきが考えられる．出産を経験するユーザは妊娠から出産まで似た過程を辿るが，内定を経験するユーザは人によって就職活動の期間などの要素が異なる．例えば，就職活動の期間が 3 ヶ月程度のユーザの場合，3 ヶ月以上前のツイートには特徴が現れにくい可能性がある．つまり，一般ユーザとは異なる特徴が出にくい期間まで遡って正例とした場合が出産の予測より多かったために，精度が下がったと考えられる．

**退院のイベントの予測** (図 4 右) は，内定と同様，出産の Precision-Recall 曲線と比較すると精度が低くなっていることがわかる．退院のイベントにおいても，ユーザごとのばらつきが原因として考えられる．退院を経験するユーザのツイートに

特徴が出やすい期間は，そのユーザが入院している期間であると考えられ，入院期間も就職活動と同様に人によって期間が異なる．また，内定の Precision-Recall 曲線と比べて退院の予測精度が低いことから，退院を経験するユーザの入院期間は平均して就職活動よりも短いと考えられる．

**妊娠のイベントの予測** (図 5 右) は，内定や退院と同様，出産の Precision-Recall 曲線と比較すると精度が低くなっている．理由として，Twitter 上での妊娠報告を，妊娠発覚後どの段階で行うかが人によって異なることが挙げられる．妊娠を経験したユーザがイベントの発生に言及したツイートを確認したところ，妊娠 3 ヶ月目の段階で報告しているユーザから，遅い場合は 8 ヶ月目の段階で報告しているユーザまで存在し，報告時期にばらつきが大きかった．同じ妊娠イベントの直前 1 ヶ月のツイート集合でも，それが妊娠 3 ヶ月目の段階であったり，妊娠 8 ヶ月目の段階であったりするため，$K = 6$ としたときに，一般ユーザとは異なる特徴が出にくい期間まで遡って正例とした結果，精度が低下したと考えられる．

**結婚のイベントの予測** (図 6 右) は，本研究で対象としたライフイベントの中で最も予測精度が低い．また表 3 より，AUC も対象としたライフイベントの中で最も低い．このことから，結婚は他のイベントに比べると，イベントユーザと一般ユーザのツイートに現れる単語の特徴にあまり差がないことがわかる．また，$K$ の値を小さくすると AUC の値が低下する傾向にあることから，結婚ではイベント開始直前になってもイベントユーザのツイートにイベントに特有な単語が現れにくいことがわかる．実際に各期間におけるイベントユーザのツイート集合中に出現する単語を確認したところ，いずれの期間においても，他のイベントと比較して結婚に関連する特徴的な単語の数が少なかった．以上の結果から，結婚は他のイベントと異なる性質を持ち，提案手法のような近い将来にイベントを経験するユーザが使用する単語の特徴を用いる手法では予測することが難しいイベントであると考えられる．このようなイベントに対しては，単語の出現傾向だけでなく他の要素を考慮しなければならない．例えば，各期間におけるイベントユーザのツイート数の変化や，他のユーザとの交流パターンの変化などが考えられる．このような特徴を考慮することで結婚のようなイベントの発生を予測可能になると考えられる．

## 4. ま と め

本研究では，ライフイベントを近い将来に経験する Twitter ユーザの予測を目的として，ユーザのツイート履歴に基づくライフイベント予測モデルの構築手法を提案した．評価実験では，既存手法により抽出できるライフイベントの中で，多くの人が生涯の中で経験すると思われる「出産」,「内定」,「退院」,「妊娠」,「結婚」の 5 種類のイベントを評価の対象とした．その結果，出産，内定，妊娠，結婚のイベントにおいては，予測期間 $K$ を 6 と大きくした際に，提案手法がキーワードに基づくベースライン手法よりも精度良くライフイベントを予測できた．しかし，$K$ を 1 として予測した場合は，1 ヶ月以内にライフイベントを経験するユーザと数ヶ月後にライフイベントを経験するユーザを判別することが難しく、キーワードを用いたベースライン手法の性能を上回ることができなかった。

本研究では，分類器を学習する際の特徴として単語の出現確率を用いていたが，結婚など，単語として特徴が現れにくいライフイベントにおいては精度が低くなる傾向があった．しかし，単語の情報以外にも，ライフイベントの予測に有効な様々な特徴が考えられる．例えば，ユーザのプロフィールや他のユーザとの交流パターンなどを用いて，ライフイベント発生前の特有の傾向を捉えられる可能性がある．また本研究では，予測期間 $K$ を設定し，正例，負例のラベル付けを行っていたが，多くのイベントにおいてユーザごとに特徴が出やすい期間が異なっていた．そのため，一律に期間を設定することがライフイベント予測の問題において妥当であるかどうかについても検討する必要がある．加えて，イベントの発生が近づくにつれ，ユーザが使用する単語やツイート数，他のユーザとの交流パターンが変化することが考えられるため，時系列の情報を扱えるモデルを構築することを検討している．

**謝辞**　本研究の一部は，文部科学省科学研究費補助金・基盤研究 (A)(26240013)，JST 国際科学技術共同研究推進事業（戦略的国際共同研究プログラム），および，KDDI 総合研究所の研究助成によるものである．ここに記して謝意を表す．

### 文　献

[1] T. Sakaki, M. Okazaki and Y. Matsuo: “Earthquake shakes twitter users: real-time event detection by social sensors”, Proc. World Wide Web Conf. (WWW), pp. 851–860 (2010).
[2] J. Li and C. Cardie: “Early stage influenza detection from twitter”, arXiv preprint arXiv:1309.7340 (2013).
[3] B. Di Eugenio, N. Green and R. Subba: “Detecting life events in feeds from twitter”, Proc. IEEE Int’l Conf. on Semantic Computing (ICSC), pp. 274–277 (2013).
[4] S. Choudhury and H. Alani: “Personal life event detection from social media”, Proc. Workshop on Social Personalisation (SP) (2014).
[5] T. Dickinson, M. Fernandez, L. A. Thomas, P. Mulholland, P. Briggs and H. Alani: “Identifying prominent life events on twitter”, Proc. Int’l Conf. on Knowledge Capture (K-CAP) (2015).
[6] L. G. Moyano, P. R. Cavalin and P. P. Miranda: “Life event detection using conversations from social media”, Proc. Brazilian Workshop on Social Network Analysis and Mining (BraSNAM) (2015).
[7] J. Li, A. Ritter, C. Cardie and E. H. Hovy: “Major life event extraction from twitter based on congratulations/condolences speech acts”, Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp. 1997–2007 (2014).
[8] A. Hurriyetoglu, N. Oostdijk and A. van den Bosch: “Estimating time to event from tweets using temporal expressions”, Proc. Workshop on Language Analysis for Social Media (LASM) (2014).
[9] H. Becker, D. Iter, M. Naaman and L. Gravano: “Identifying content for planned events across social media sites”, Proc. Int’l Conf. on Web Search and Data Mining (WSDM), pp. 533–542 (2012).