Vol.2018-DBS-167 No.7
Vol.2018-IFAT-132 No.7
2018/9/12

# URLエンベディングを用いたライフイベント予測

星　尚志[1,a)]　秋山 卓也[1,b)]　木村 塁[1,c)]　黒柳　茂[2,d)]　南川　敦宣[1,e)]

**概要**：近年，インターネットが身近にあり，Web アクセスを介して購買行動や，調査，SNS などに利用されている．また Web アクセス解析を実施することでユーザの行動を理解し効率的な広告配信などのマーケティングに利用する技術も存在する．アクセス解析は Web アクセスログからユーザのアクセス傾向を特徴量化し，ディープラーニング等を利用してモデル化することでユーザの趣向を識別することができる技術である．しかしながら Web アクセスログは膨大であり，ユーザ一人ひとりについて分析することは時間，処理量ともに限界がある．そこで我々は Web アクセス傾向の特徴量について次元圧縮を行うことで処理量を削減したアクセス解析の手法を提案する．ユーザの Web アクセスログについて，URL ドメインを単語，アクセス遷移を文章とみなした上で word2vec により次元圧縮を行う．評価として 2016 年に実施したアンケート回答ユーザの回答結果と Web アクセスログを利用し，転居，結婚等のライフイベントを予測するモデルについて，既存の手法との精度比較を行った．

# Life event prediction using URL embedding

Hisashi Hoshi[1,a)]　Takuya Akiyama[1,b)]　Rui Kimura[1,c)]　Shigeru Kuroyanagi[2,d)]
Atsunori Minamikawa[1,e)]

## 1. はじめに

近年，インターネットの利用者は，検索エンジンを通じた情報検索や E コマースウェブサイトを通じたショッピングなどを利用する過程で，自ら能動的にサイトを閲覧するだけでなく，広告レコメンドシステムから受動的に情報を受け取る．これらのシステムから配信された広告がユーザによって期待される情報である場合，クリック率やコンバージョン率が向上すると考えられており，ユーザに適した広告を配信するために，広告レコメンドシステムは，予め誰がターゲットユーザであるかを推定する必要がある．また，広告レコメンドシステムを運営する広告配信事業者は，広告主に対し，性別や年齢などのユーザー属性・アクセス元から推定したエリア・ユーザーの興味や，関心・ライフイベントやライフステージといったターゲティング条件を提供するのが一般的であり，そのターゲティング条件の種類やターゲティングによって配信可能なユーザー数が多いことが広告配信事業者の競争力の一つとなっている．ターゲティング条件の生成には，媒体側がユーザーから取得した属性情報をオーディエンスデータとして直接取得する方法の他，ユーザーの Web 上での行動を利用してユーザー属性やライフイベントなどを推定する手法が取られるのが一般的となっている．

ユーザのターゲティング条件の中では，結婚や引っ越しなどのライフイベントは消費者の購買行動に非常に影響し，大きな金額の消費が発生するため，ターゲティング条件としての価値も高い．ライフイベントに直面しているユーザは，事前に関連情報を検索し，関連する製品やサービスを購入する傾向がある．例えば，引っ越しをするユーザは，事前に賃貸住宅に関する情報を調査したり，家具を注文する傾向があり，一方，新婚者は結婚式を実施する会場や関

[1] 株式会社 KDDI 総合研究所
2-1-15, Oohara, Fujimino-city, Saitama, 356-8502, Japan
[2] Supership 株式会社
5-4-35, Minami-Aoyama, Minato-ku, Tokyo, 107-0062, Japan
[a)] hi-hoshi@kddi-research.jp
[b)] ta-akiyama@kddi-research.jp
[c)] ui-kimura@kddi-research.jp
[d)] shigeru.kuroyanagi@supership.jp
[e)] at-minamikawa@kddi-research.jp

連サイト等の情報を調査したり，ウェディングドレスを購入する傾向がある．近い将来にユーザのライフイベントの発生が分かっている場合，広告推薦システムは関連する広告を正確に配信することが可能である．

しかし，ユーザのライフイベントはプライバシー情報であり，オーディエンスデータなどを通じて直接取得することは困難である．そこで電子商取引に関するユーザのデータに基づく分析技術が最近の研究で研究されている．[1] では，最大エントロピーセミマルコフモデルを用いて，中国最大の E コマースウェブサイト，淘宝網のユーザの買い物履歴を利用して，母と子の状態を分析している．[2] では，Alibaba のデータサイエンスと機械学習の競争ウェブサイトである Tianchi のショッピング履歴の公開データを使用して，母と子の状態に関する予測および製品推奨アプローチを検討している．Frey ら [3] は購買ログが取得できない場合に，スマートフォンにインストールしたアプリに基づいて，個人の現在のライフステージを予測しており，安倍ら [4] は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から SVM によってライフイベントの予測モデルを構築する手法を提案している．Yang ら [5] は，直接ライフイベントやライフステージの推定は行っていないが，E コマースサイトでの商品クリック履歴を LSTM を用いて分散表現にすることでライフステージのモデリングを行い，翌日に購入される商品の予測を行っている．

また，Web アクセス履歴に基づいてユーザのライフイベントを予測する試みも行われている [6], [7], [8]．[6], [7] では，ユーザの Web アクセスデータからキーワードを抽出し，あらかじめ定義されたライフイベント関連キーワードと比較することでライフイベントの発生を予測する．[8] の予測アプローチでは，キーワードの代わりに Web サイト (URL) が利用されることを除いて同様である．これらの研究では，ライフイベントに関連するキーワードやウェブサイトの定義済みデータベースが必要であるが，この種のデータベースの生成方法については説明されていない．

しかしながら Web アクセスログは膨大であり，ユーザ一人ひとりについて Web アクセスログを分析することは時間，処理量ともに限界がある．また一般的な手法では，同一ドメインにアクセスしないと特徴量として有効な表現が不可能であった問題点に対し，我々は Web アクセス傾向の特徴について次元圧縮・ベクトル化を行うことで，類似する Web サイトドメインを表現可能なアクセス解析・ライフイベント予測の手法を提案する．ライフイベントの変化がある時，その前後の期間には類似した Web サイトを連続して訪問すると考えられる．そこでユーザの Web アクセスログについて，URL ドメインを単語，一連のアクセス遷移を文章とみなした上で word2vec により次元圧縮 (URL エンベディング) を実施することで，Web アクセスの特徴を保持した次元圧縮が可能となる．

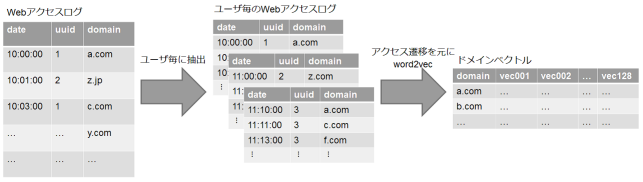


**図 1** ドメインのベクトル化

本提案手法では，以下の特徴を持つ．

- 予測対象が変わっても共通の特徴ベクトルを利用するため，データの再利用が可能となり，効率的に大量のプロファイルを予測する事が可能
- ベクトルの次元を指定することが可能であるため，低次元の特徴ベクトルが生成可能
- Web サイトの内容を参照せずにユーザの Web アクセス遷移のみで効果的な特徴量を生成することが可能

本方式の評価によって，低次元で既存手法と同等の結果が得られた．さらに複数の特徴量を組み合わせることでより高い精度が得られた．

## 2. 関連研究

Kanagasabai ら [9] は，通信事業者が保有するアクセスログに記録された各サイト URL を文章とみなし，word2vec を用いてサイトのカテゴライズを行っている．田頭ら [10] は単一ウェブサイトのアクセス履歴に対し，各 URL を単語，ユーザーを文章とみなし，Distributed Memory Model of Paragraph Vector を用いてユーザーをベクトル化し，広告クリック及びサイト訪問の予測を行っている．Niu ら [11] は，ライフイベントに関連するサイトへのアクセスの変動を元にライフイベントを推定する手法を提案している．Wang ら [12] は，特定サイト内のユーザーの行動をイベント種別ごとにカテゴリ分けし，その行動を分析することにより，ユーザーの行動予測が可能な事を示している．

## 3. 提案手法

本章では提案手法の説明を行う．結婚や引っ越しのようなライフイベントの変化は頻繁に発生することはなく，かつ人生の中で大きな変化である．そのため，ライフイベントの変化前はそれらの内容について Web サイト等で情報検索を行うことが考えられる．またさまざまな情報を収集するため，短期間に類似した Web サイトを多く閲覧することが考えられる．そこで，本研究では Web サイトのアクセス傾向について，URL ドメインとアクセス遷移を利用した word2vec を実施することにより，類似したカテゴリの Web サイトを距離の近いベクトルで表現可能な手法を提案する．また，ベクトルの次元を指定することが可能であるため，低次元の特徴ベクトルが生成可能となる．本手法の概要図を図 1，2 に示す．

表 1　任意のドメインと類似するドメイン

| ドメイン | 子ども命名サイト A | Cos 類似度 | 旅行サイト A | Cos 類似度 | ポイントサイト A | Cos 類似度 | レシピサイト A | Cos 類似度 |
|---|---|---|---|---|---|---|---|---|
| 1 | 子ども命名サイト B | 0.973 | 旅行サイト B | 0.891 | ポイントサイト B | 0.849 | グルメサイト B | 0.724 |
| 2 | 子ども命名サイト C | 0.971 | 旅行サイト C | 0.812 | ポイントサイト C | 0.840 | 女性関連サイト | 0.678 |
| 3 | 子ども命名サイト D | 0.970 | 旅行サイト D | 0.811 | ポイントサイト D | 0.751 | カメラ関連サイト | 0.641 |
| 4 | 子ども名前検索サイト A | 0.967 | 旅行サイト E | 0.810 | ポイントサイト E | 0.714 | レシピサイト C | 0.608 |
| 5 | 子ども名前検索サイト B | 0.966 | 旅行サイト F | 0.808 | ポイントサイト F | 0.688 | ブログ | 0.580 |
| 6 | 姓名判断サイト | 0.953 | 旅行サイト G | 0.803 | ISP サイト | 0.667 | レシピサイト D | 0.579 |
| 7 | 子ども命名サイト E | 0.952 | 旅行サイト H | 0.802 | ポイントサイト G | 0.656 | レシピサイト E | 0.578 |
| 8 | 子ども命名サイト F | 0.949 | 観光サイト A | 0.795 | 不明 | 0.652 | レシピサイト F | 0.578 |
| 9 | 子ども命名サイト G | 0.940 | 観光サイト B | 0.794 | 小遣い稼ぎサイト A | 0.647 | レシピブログ | 0.565 |
| 10 | 子ども命名サイト H | 0.937 | 旅行サイト I | 0.791 | 小遣い稼ぎサイト B | 0.639 | レシピサイト G | 0.561 |

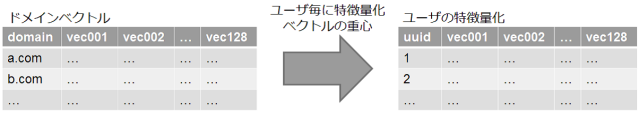


図 2　ユーザの特徴量化

### 3.1　アクセスログ抽出

本研究では Supership 株式会社が保有しているビッドリクエストログを利用する．ビットリクエストとは，ユーザが媒体社のサイトに訪れた際に SSP(Supply Side Platform) が DSP(Demand Side Platform) に広告配信枠に広告を配信するかを決めるための入札リクエストログのことであり，cookie 単位で保持される．ビッドリクエストログには媒体社の URL が含まれており，本研究ではビッドリクエスト毎に URL からドメインを抽出しておく．このビッドリクエストログについて，ある一定の期間におけるビッドリクエストログを収集し，cookie 単位で URL ドメインの抽出を行う．

### 3.2　URL エンベディング

低次元の特徴量を生成する．アクセス遷移の特徴量生成手法として，word2vec[13] を利用した．cookie 単位で抽出したビッドリクエストログに対して，図 1 に示すように word2vec を実施し 128 次元に次元圧縮することで，ドメイン毎の特徴量ベクトルを生成する．最後に図 2 に示すように，cookie 毎に生成された特徴量ベクトルの重心を算出することで最終的なユーザ毎の特徴量とする．

### 3.3　事前検証

本節では事前検証として，ライフイベントの変化のあったユーザの Web アクセスログに対して，URL ドメインを抽出し word2vec にて URL エンベディングを実施することで類似するカテゴリのドメインを類似するベクトルで表現することが可能か検証を行った．例えば出産を控えているユーザは，子育てや命名などに関する Web ページに高頻度にアクセスすることが考えられる．

Web アクセスログに対して URL エンベディングを実施し，適当なドメインを 4 つ選択してこれらのドメインベクトルとコサイン類似度が最も高いドメインベクトルを算出した．選択したドメインのカテゴリは子ども命名サイト関連，旅行サイト関連，ポイントサイト関連，レシピサイト関連である．それぞれのドメインに対するコサイン類似度の上位 10 件を表 1 に示す．

表 1 より，子どもの命名に関する Web サイトのドメインである「子ども命名サイト A」は，0.9 以上のコサイン類似度で他の命名サイトに関するドメインを抽出することができた．また旅行関連，ポイントサイト関連についても高いコサイン類似度でドメインを抽出することができている．レシピ関連 Web サイトであるドメイン「レシピサイト A」は他の 3 ドメインと比較してコサイン類似度は低いが，グルメサイトや女性に関する Web サイトが上位に抽出されており，ユーザの嗜好を表現することができていると考えられる．

## 4.　評価

本章では 2016 年 9 月に実施したアンケート回答者 6,592 名を対象に，提案手法にて説明した手法を利用したライフイベント予測とその評価結果について説明する．

### 4.1　評価方法

予測する対象は，結婚，転居，出産，転職のライフイベントとし，特徴量の生成には，各ライフイベントの 2 ヶ月前から当月までの 3 ヶ月間の Web アクセスログを利用した．評価対象のユーザは，上記 3 ヶ月間に 5 つ以上のドメインにアクセスし，かつアクセス数上位 128 ドメインに 1 ドメイン以上アクセスしたユーザとした．そのため各ライフイベント予測対象は表 2 に示すようになる．

word2vec による URL エンベディングには 2016/04/16 から 2016/07/10 までの 1,943,755 名の Web アクセスログを利用し，出現するドメインを 128 次元でベクトル化した．

提案手法は，3 ヶ月間の Web アクセスログに出現するドメインに対し，各ユーザ毎に上記 word2vec 特徴量ベクトルの重心を算出して，128 次元の特徴量ベクトルを生成した．

比較対象はアクセスドメイン数の上位 128 ドメインへの

**表 2** 予測対象と正例の出現率

| 予測対象 | UU 数 [人] | 正例 [%] |
|---|---|---|
| 結婚 | 2084 | 0.96 |
| 転居 | 1920 | 4.15 |
| 出産 | 1399 | 6.51 |
| 転職 | 1969 | 6.30 |

**表 3** ライフイベント予測結果

| 予測対象 | 特徴 | Precision [%] | Recall [%] | Lift | AUC |
|---|---|---|---|---|---|
| 結婚 | アクセス数上位 128 | 0.0 | 0.0 | 0.0 | 0.652 |
| | オッズ比上位 128 | 0.0 | 0.0 | 0.0 | 0.595 |
| | 提案手法 | 0.0 | 0.0 | 0.0 | **0.710** |
| 転居 | アクセス数上位 128 | 8.1 | 3.4 | 1.24 | **0.561** |
| | オッズ比上位 128 | 7.0 | 3.4 | 1.08 | 0.515 |
| | 提案手法 | 19.1 | 1.0 | 2.94 | 0.549 |
| 出産 | アクセス数上位 128 | 5.6 | 2.2 | 1.34 | 0.642 |
| | オッズ比上位 128 | 18.0 | 7.8 | 4.33 | 0.667 |
| | 提案手法 | 7.0 | 0.3 | 1.69 | **0.699** |
| 転職 | アクセス数上位 128 | 13.7 | 5.2 | 2.18 | **0.595** |
| | オッズ比上位 128 | 11.5 | 4.2 | 1.82 | 0.542 |
| | 提案手法 | 2.4 | 0.2 | 0.386 | 0.585 |

**表 4** 複数特徴を組み合わせたライフイベント予測結果

| 予測対象 | 特徴 | Precision [%] | Recall [%] | Lift | AUC |
|---|---|---|---|---|---|
| 結婚 | 提案手法+アクセス数上位 128 | 0.0 | 0.0 | 0.0 | 0.707 |
| | 提案手法+オッズ比上位 128 | 0.0 | 0.0 | 0.0 | **0.713** |
| 転居 | 提案手法+アクセス数上位 128 | 17.5 | 1.0 | 2.68 | 0.582 |
| | 提案手法+オッズ比上位 128 | 19.2 | 1.0 | 2.95 | **0.583** |
| 出産 | 提案手法+アクセス数上位 128 | 23.3 | 1.2 | 5.63 | 0.710 |
| | 提案手法+オッズ比上位 128 | 4.2 | 0.3 | 1.01 | **0.738** |
| 転職 | 提案手法+アクセス数上位 128 | 6.9 | 0.5 | 1.09 | **0.601** |
| | 提案手法+オッズ比上位 128 | 4.8 | 0.3 | 0.754 | 0.599 |

3 ヶ月間のアクセス回数，及びオッズ比上位 128 ドメインへの 3 ヶ月間のアクセス回数を特徴量としたものを用い，それぞれ 128 次元の特徴量となっている．オッズ比上位 128 ドメインについては，各予測対象毎にオッズ比を求めて上位 128 ドメインを対象としているため，予測対象毎に対象となるドメインが異なる．分類器は全て共通で XGBoost を利用した．結果のばらつきを考慮し，5-fold 交差検定を 10 回実施し，ROC-AUC で評価を実施した．

### 4.2 評価結果

ライフイベント予測結果を表 3 に示す．表 3 より結婚，出産は提案手法の AUC が最も高い結果となった．しかしながらその他の推定対象については他の手法と比較して低い AUC となった．そこで提案手法とアクセス数上位 128，オッズ比上位 128 を組み合わせた特徴量で再学習し，評価を実施した．結果を表 4 に示す．表 4 より，提案手法に複数手法を組み合わせる事により AUC が向上されることが示された．

## 5. 考察

アクセス数上位 128 ドメインやオッズ比上位 128 ドメインの特徴量は，同一のドメインにアクセスしないと特徴量として有効な表現が不可能だった．しかし提案手法では，類似したカテゴリの Web サイトを距離の近いベクトルで表現可能であるため，ライフイベント予測では比較した手法よりも精度向上に寄与したと考えられる．

## 6. おわりに

ユーザの Web アクセスログを用いたライフイベント予測に関して，一般的な手法は同一ドメインにアクセスしないと特徴量として有効な表現が不可能であった問題点に対し，Web アクセスログに関してエンベディングを実施することで解決する手法を提案した．これは Web サイトのアクセス傾向について，URL ドメインとアクセス遷移を利用した word2vec を実施することで類似したカテゴリの Web サイトを距離の近いベクトルで表現可能な手法であり，かつ低次元の特徴ベクトルが生成を可能とした．本手法と既存手法に関して特徴量の次元を揃えた評価の結果，低次元で既存手法と同等の結果が得られた．さらにアクセス数上位 128 ドメイン，オッズ比上位 128 ドメインと提案手法の特徴量を組み合わせることでより高い精度が得られ，本手法の有効性が示された．今後は Web アクセスの順序を考慮したモデルや，Web アクセス間隔を考慮したモデル等，より精度の高い手法を検討する予定である．

### 参考文献

[1] Peng Jiang, Yadong Zhu, Yi Zhang and Quan Yuan, “Life-stage Prediction for Product Recommendation in E-commerce”, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pages 1879-1888, August 10-13, 2015, Sydney, NSW, Australia.
[2] Bin Guo, Kai Dou and Li Kuang, “Life Stage Based Recommendation in E-commerce”,2016 International Joint Conference on Neural Networks, Pages 3461-3468, July 24-29, 2016, Vancouver, BC, Canada.
[3] Remo Manuel Frey, Runhua Xu and Alexander Ilic, “Mobile App Adoption in Different Life Stages: An Empirical Analysis”, Pervasive and Mobile Computing, vol. 40, Pages 512-527, Sept. 2017.
[4] Shun Abe, Masumi Shirakawa, Takahiro Hara,Kazushi Ikeda and Keiichiro Hoashi, “Construction of Life Event Prediction Model using Tendency of Word Occurrence in User’s Tweet History”,IEICE technical report, vol. 117, no. 108, Pages 1-6, Jun. 2017.
[5] Jing-Wen Yang, Yang Yu and Xiao-Peng Zhang, “Life-stage modeling by customer-manifold embedding”, In Proceedings of the 26th International Joint Conference on Artificial Intelligence (pp. 3259-3265).
[6] NTT and Tokyo Institute of Technology, unexamined patent application 2011-227746, 2011-11-10.
[7] NTT and Tokyo Institute of Technology, unexamined patent application 2013-125495, 2013-06-24.
[8] Dai Nippon Printing Co., Ltd, unexamined patent application 2017-117351, 2017-06-29.
[9] Kanagasabai, Rajaraman, et al. ”Classification of massive mobile web log URLs for customer profiling &

alytics.” Big Data (Big Data), 2016 IEEE International Conference on. IEEE, 2016.
[10] 田頭幸浩, et al. ”オンライン広告におけるウェブ閲覧系列の分散表現の獲得.” 人工知能学会全国大会論文集 2016 年度人工知能学会全国大会 (第 30 回) 論文集. 一般社団法人 人工知能学会, 2016.
[11] Hao Niu, Mori Kurokawa, Shigeru Kuroyanagi and Arei Kobayashi, “Mining Life Events Based on the Fluctuation of Users’ Web Access”, 第 10 回 Web とデータベースに関するフォーラム (WebDB Forum 2017).
[12] Gang Wang, Xinyi Zhang, Shiliang Tang, Haitao Zheng and Ben Y. Zhao, “Unsupervised Clickstream Clustering for User Behavior Analysis”, In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 225-236). ACM.
[13] word2vec, `https://code.google.com/p/word2vec/` (2018/08/08 アクセス)

# 正誤表

下記の箇所に誤りがございました．お詫びして訂正いたします．

| 訂正箇所 | 誤 | 正 |
|---|---|---|
| 概要<br>9 行目 | Web アクセスログ | ビッドリクエストログ |
| 3 ページ<br>3.1 節 2 行目 | ビットリクエスト | ビッドリクエスト |
| 3 ページ<br>3.3 節，4.1 節<br><br>4 ページ<br>6 章 4 行目 | Web アクセスログ | ビッドリクエストログ |