情報処理学会論文誌　データベース　Vol.10　No.4　6–10 (Dec. 2017)

テクニカルノート

非負値行列分解を用いた時系列リンク予測

中嶋 篤宏[1,a)]　佐々木 勇和[1,b)]　鬼塚 真[1,c)]

受付日 2017年6月8日, 採録日 2017年8月2日

概要：ソーシャルネットワークやWebページの遷移などに見られるように，多くのデータがグラフ構造として表されており，様々なデータマイニングタスクに対してグラフ構造を利用した手法が研究されている．本論文では，時間的に変化するグラフ構造のリンクを予測する問題を対象とする．従来手法では，時間的順序を持つ複数のグラフを単一のグラフに統合しリンク予測を行う手法や，2次元の行列で表されるグラフを時間方向に拡張した3次元テンソルに対してテンソル分解を行いリンクを予測する手法が研究されてきたが，時間方向の連続的な変化をとらえることが難しいため予測精度が高くないという問題があった．提案手法では，時間的順序を持つ複数のグラフに対して非負値行列分解を用いて対象データの特徴を低次元に集約し，Holt-Winters予測を用いて季節変動やトレンドなどの周期的な変化をとらえた予測を行う．実データを用いて比較実験を行い，提案手法は従来手法より高い性能を示すことを確認した．

キーワード：リンク予測，非負値行列分解，Holt-Winters予測

Temporal Link Prediction by Non-negative Matrix Factorization

Atsuhiro Nakashima[1,a)]　Yuya Sasaki[1,b)]　Makoto Onizuka[1,c)]

Received: June 8, 2017, Accepted: August 2, 2017

Abstract: As seen in the social networks and web pages, many data are expressed in graph structures and many researches for graphs have been conducted for various data mining tasks. In this article, we focus on the temporal link prediction problem for bipartite graphs. The typical techniques for the problem predict temporal links by either merging a temporal graph sequence into a single static graph or by applying tensor factorization after transforming the temporal graph sequence into a single tensor with additional temporal axis. However, the prediction accuracy for those techniques is not high because they have a difficulty in capturing continuous changes in the time axis. We design an algorithm that combines non-negative matrix factorization and Holt-Winters in order to extract latent features from bipartite graphs and to predict periodic changes in future. We conduct experiments using a real data set and demonstrate that our algorithm achieves higher performance over existing techniques.

Keywords: link prediction, non-negative matrix factorization, Holt-Winters prediction

1. はじめに

ソーシャルネットワークやWeb解析，協調フィルタリングなどの様々なアプリケーションのデータは，リンクで表されるオブジェクト間の関係を含んでいる．たとえば，2人の人間が互いに連絡先を交換しているならば，その関係をリンクで表すことができる．これらの関係は，グラフ構造でモデル化でき，オブジェクトを表すノードとリンクを表すエッジで表現される．このようにグラフを用いることで，隠れたグループの検出や欠落したリンクの予測，オブジェクトのランク付けなど様々なタスクに応用することができる [3], [6]．ソーシャルネットワークやWebページのリンクなどの実世界のデータは，時間とともに変化するため，時間的変化を含むデータに対してリンクがどのように変化するか予測することが必要である [5], [9]．つまり，時刻1から時刻Tまでのグラフ構造が与えられたとき，次の時刻$T+1$のグラフ構造を予測するという課題である．

[1] 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University, Suita, Osaka 565–0871, Japan
[a)] nakashima.atsuhiro@ist.osaka-u.ac.jp
[b)] sasaki@ist.osaka-u.ac.jp
[c)] onizuka@ist.osaka-u.ac.jp

© 2017 Information Processing Society of Japan

この問題に対して，今まで様々な時系列リンク予測手法が提案されてきたが，時間方向の連続的な変化をとらえることが難しく予測精度が高くないという問題がある [2]．この課題を解決するためには，グラフの特徴を適切にモデル化すること，その特徴の周期的な変化をとらえることが必要である．グラフの特徴をモデル化する手法として非負値行列分解が考えられる．非負値行列分解は，多くの利用者を対象とした行動を分析する際に，共通する特徴を抽出して行動をモデル化する有効な手段として利用されている [8], [11]．一方，データの周期的な変化をとらえることのできる代表的な手法として Holt-Winters 法がある [1], [7]．Holt-Winters 法を用いることで，実世界のデータに含まれる季節変動や，オリンピックなどの定期的なイベントの影響による変化のような周期的な変化をとらえた予測が可能になる．

本論文ではこれらを組み合わせることで，複数利用者の行動に共通する特徴を抽出すると同時に，これらの特徴に関する周期性モデルをとらえることで，行動を予測する手法を提案する．提案手法と従来手法を実データを用いて比較した結果，提案手法が高い予測性能を持ち，特に潜在的なリンクの予測に対して高い性能を持つことを確認した．

本論文の構成は以下のとおりである．2 章で関連研究を述べる．3 章で提案手法の説明に必要な事前知識を述べる．4 章で提案手法の詳細を説明する．5 章で評価実験の結果を示す．6 章で結論を述べる．

2. 関連研究

時間的に変化するリンクの予測問題に対して様々な研究が行われている．この章では，時系列リンク予測を行う様々な手法について述べる．

グラフ構造は，隣接行列と呼ばれる 2 次元の行列で表すことができる．あるノード i と j の間にエッジがあるとき，エッジの本数，もしくはエッジの重みを行列の (i, j) 成分に割り当てる．エッジを持たない場合は，0 を割り当てることで表現できる．このようにして，N 個のノードを持つグラフであれば，$N \times N$ の行列として表現でき，2 つのノード集合に分かれるような 2 部グラフであれば，$M \times N$ の行列として表現できる．時間的に変化するグラフ構造を表現するためには，時間方向の軸を追加した 3 次元のテンソルが必要である．

行列ベースの時系列リンク予測手法として Truncated SVD（T-SVD）を用いた手法がある [4]．T-SVD は行列の低次元近似法であり，行列を 2 つの直行行列と 1 つの特異値行列に分解し，値の大きい K 個の特異値を使って，元の行列を K ランクの 3 つの行列で近似する．T-SVD を用いて時系列リンク予測を行うために，CT，CWT と呼ばれる手法を用いて 3 次元テンソルを 2 次元の行列に次元削減する [2]．この行列に対して T-SVD を適用することで，その行列の特徴を抽象化した行列を求めることができ，この特徴抽象化行列を用いてリンク予測を行う．

また，テンソルベースの時系列リンク予測手法として，CANDECOMP/PARAFAC（CP）分解がある [2], [10]．CP 分解は，1 つの 3 次元テンソルを 3 つの 2 次元行列に分解する．3 つの行列は，元の 3 次元テンソルの各軸に対応しており，時間軸に対応する行列の時間的に新しい一定時間の平均を用いて，新しいリンク情報がより大きな影響を持つようにリンク予測を行う．

3. 事前知識

まず，3.1 節で非負値行列分解について述べ，3.2 節で Holt-Winters 法について述べる．

3.1 非負値行列分解

グラフ構造を表す隣接行列では，値はリンクの有無，もしくは重みの値であり，負の値を含まない自然数で表される．非負値行列分解（NMF: Non-negative Matrix Factorization）は，このような非負の値を持つ行列を低次元の 2 つの行列に分解する手法である．NMF は減算を行わないという制約を持つため，元の行列に負の値が含まれない場合は，分解行列の値にも負の値が含まれない．この制約により，NMF は特異値分解などの他の行列分解手法とは異なる分解結果をもたらし，元の行列内のいくつかの頻出パターンを抽出することができる．その抽出された頻出パターンに基づいて，データの要素がクラスタリングされ，それらの分解結果を推薦システムなどのタスクに利用することができる．

NMF では，$I \times J$ サイズの非負値行列 $\mathbf{X}$ を $I \times K$ サイズの非負値行列 $\mathbf{U}$ と，$K \times J$ サイズの非負値行列 $\mathbf{V}$ に分解する．K は NMF の基底数であり，任意のパラメータである．本論文では，乗法的更新アルゴリズムを用いて，分解行列を求める．乗法的更新では，ランダムな非負値で初期化した行列 $\mathbf{U}$，$\mathbf{V}$ に対して，元の行列 $\mathbf{X}$ と分解行列の積 $\mathbf{UV}$ のユークリッド距離を計算し，以下の更新式に従って繰り返し計算を行うことにより，その距離を最小化することで最終的な分解行列 $\mathbf{U}$，$\mathbf{V}$ を求める．

$$v_{kj}^{(t+1)} = v_{kj}^{(t)} \frac{\sum_i u_{ik}^{(t)} x_{ij}}{\sum_i u_{ik}^{(t)} \sum_k u_{ik}^{(t)} v_{kj}^{(t)}} \tag{1}$$

$$u_{ik}^{(t+1)} = u_{ik}^{(t)} \frac{\sum_j v_{kj}^{(t+1)} x_{ij}}{\sum_j v_{kj}^{(t+1)} \sum_k u_{ik}^{(t)} v_{kj}^{(t+1)}} \tag{2}$$

この更新計算を繰り返し，得られた分解行列は SVD 同様，元の行列の特徴を低次元に集約した行列と見なすことができる．つまり，行列の各軸の特徴を基底 K の数のグループに集約しているといえる．

© 2017 Information Processing Society of Japan

3.2 Holt-Winters 法

Holt-Winters 法は，季節性やトレンドなどの周期的な時系列データに適した予測手法である．Holt-Winters 法は，以下に示す，1 つの予測方程式と 3 つの指数方程式で構成される．

$$y_{t+h} = l_t + hb_t + s_{t-m+h}. \tag{3}$$

$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}). \tag{4}$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}. \tag{5}$$

$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}. \tag{6}$$

$y_1, y_2, \ldots, y_t$ を観測値とすると，y_{t+h} は h 時間先の予測値である．m は周期を表す任意のウィンドウサイズであり，α，β，γ は指数係数である．式 (4) は，季節性を除いた観測値と非季節性成分の加重平均である．式 (5) は，1 単位時間前の増減量の加重平均である．式 (6) は，現在と 1 周期前の季節性成分の加重平均である．式 (4)–(6) によって，時間 t の水平成分 l_t，傾向成分 b_t，季節性成分 s_t を求める．この 3 つの指数方程式を用いて 1 単位時間先の誤差の二乗を最小化し，予測方程式 (3) によって周期性をとらえた予測値を計算する．

4. 提案手法

提案手法は，NMF を用いて，時間ごとに複数利用者の行動に共通する特徴を抽出し，Holt-Winters 法によって，抽出した特徴の周期的な変化をとらえ，未来の行動を予測する．**図 1** に手法の概要図を示す．

提案手法ではまず，時間軸でデータを T 個に分割し，分割したデータ $\mathbf{X^{(1)}}, \ldots, \mathbf{X^{(T)}}$ に対して NMF を適用することで，分解行列 $\mathbf{U^{(1)}}, \mathbf{V^{(1)}}, \ldots, \mathbf{U^{(T)}}, \mathbf{V^{(T)}}$ を得る．そして，時間方向に連続する各 $\mathbf{U}$，$\mathbf{V}$ の各要素に対して Holt-Winters 法を適用し，$\mathbf{U^{(T+1)}}$，$\mathbf{V^{(T+1)}}$ を予測し，$\mathbf{X^{(T+1)}} = \mathbf{U^{(T+1)}V^{(T+1)}}$ によって最終的な予測行列 $\mathbf{S} = \mathbf{X^{(T+1)}}$ を得る．しかし，NMF では，ランダムな非負値によって分解行列を初期化するため，最終的な分解行列は初期分解行列に依存する．そのため，NMF によって行列の特徴を基底 K の数に集約できても，ランダムな初期値のために集約された特徴が各分解行列の同じ場所に出現するという保証はなく，各時間において基底ベクトル（$\mathbf{U}$ ならば行ベクトル，$\mathbf{V}$ ならば列ベクトル）の因子の順番が変わってしまうという問題がある．そこで，分割されたデータの時間方向の平均 $\mathbf{X_{Ave}}(i,j) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{X^{(t)}}(i,j)$ に対して，NMF を適用してできた分解行列を，全時間軸にわたって共通の初期分解行列 $\mathbf{U_{init}}$，$\mathbf{V_{init}}$ とする．初期分解行列を共通化することで，基底ベクトルの因子の順番が変化しにくい，つまり同じ特徴が各分解行列の同じ場所に出現しやすいようにすることができる．これにより，時間軸全体の行列の特徴をとらえることが期待できる．

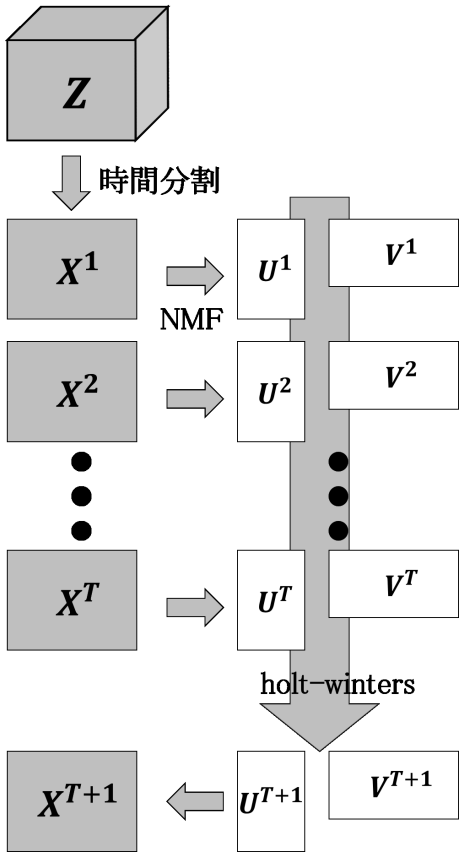


図 1　提案手法概要
Fig. 1　Overview of proposed method.

Algorithm 1 に提案手法の全体的なアルゴリズムを示す．時間軸で分割されたデータ行列のリストである $\mathbf{X}$ と分割されたデータの時間方向の平均行列 $\mathbf{X_{Ave}}$ を入力として，1 単位時間先の予測行列 $\mathbf{X^{(T+1)}}$ を出力する．平均行列にランダムな初期値 $\mathbf{U_{random}}$，$\mathbf{V_{random}}$ を用いて NMF を適用し，初期分解行列 $\mathbf{U_{init}}$，$\mathbf{V_{init}}$ を求め（1 行目），分割された各データ行列に対して初期分解行列を用いて NMF を適用し分解行列を求め（2，3 行目），分解行列のリストを作成し，そのリストを用いて Holt-Winters 法で 1 単位時間先の分解行列 $\mathbf{U^{(T+1)}}$，$\mathbf{V^{(T+1)}}$ を求め（4 行目から 8 行目），その分解行列を掛け合わせて 1 単位時間先の予測行列 $\mathbf{S} = \mathbf{X^{(T+1)}}$ を求める（9 行目）．

Algorithm 1 予測行列 $\mathbf{X^{(T+1)}}$ の計算

Input $\mathbf{X} = (\mathbf{X^{(1)}}, \mathbf{X^{(2)}}, \ldots, \mathbf{X^{(T)}}), \mathbf{X_{Ave}}$
Output $\mathbf{X^{(T+1)}}$

1: NMF : $\mathbf{U_{init}}, \mathbf{V_{init}} \leftarrow \mathbf{X_{Ave}}, \mathbf{U_{random}}, \mathbf{V_{random}}$
2: **for each** $\mathbf{X^{(t)}} \in \mathbf{X}$ **do**
3: NMF : $\mathbf{U^{(t)}}, \mathbf{V^{(t)}} \leftarrow \mathbf{X^{(t)}}, \mathbf{U_{init}}, \mathbf{V_{init}}$
4: $\mathbf{U}list \leftarrow \mathbf{U^{(t)}}$
5: $\mathbf{V}list \leftarrow \mathbf{V^{(t)}}$
6: **end for**
7: $Holt-Winters$: $\mathbf{U^{(T+1)}} \leftarrow \mathbf{U}list$
8: $Holt-Winters$: $\mathbf{V^{(T+1)}} \leftarrow \mathbf{V}list$
9: $\mathbf{X^{(T+1)}} = \mathbf{U^{(T+1)}V^{(T+1)}}$

© 2017 Information Processing Society of Japan

5. 評価実験

この章では，4 章で述べた提案手法と 2 章で述べた手法に対してリンク予測の精度を比較した評価実験と結果について述べる．

5.1 実験内容

使用したデータセットは，スーパーマーケットの 24 カ月分の POS データである．購入頻度の高いユーザ 1000 人と，購入数の多い商品 500 品を抜粋し，隣接行列を作成した．さらに，最後の 1 カ月を除く 23 カ月のデータを月ごとに分割したデータを用いて，時間方向に拡張した 3 次元テンソル Z を作成した．つまり，t の月にユーザ i が商品 j を n 個買ったとすると，$Z(i,j,t)=n$ と表される．本実験では，式 (7)–(9) に表される正規化および評価手法は文献 [2] における時系列リンク予測実験に従った．

データ内の大きい値の影響を取り除くために，

$$Z(i,j,t)=\begin{cases}1+\log(n) & n>0\\ 0 & n=0\end{cases} \tag{7}$$

として，正規化を行った．このデータを用いて，翌月のリンク情報を正解データとして予測する．正解データはユーザ i が商品 j を n 個購入したとすると，

$$\mathbf{Y}(i,j)=\begin{cases}1 & n>0\\ 0 & n=0\end{cases} \tag{8}$$

で表される行列 $\mathbf{Y}$ である．事前に決定しておくパラメータとして，Holt-Winters 法の 3 つの指数係数はそれぞれ，$\alpha=0.3$，$\beta=0.1$，$\gamma=0.1$ とし，周期は 3 カ月とした．これは，販売情報のようなデータには 4 半期ごとに季節性が現れると考えたためである．比較手法として，2 章で説明した 3 つの手法，T-SVD CT，T-SVD CWT，CP に加えて，$\mathbf{X_1},\ldots,\mathbf{X_T}$ から Holt-Winters 法で各要素を予測する手法 HW，CP 分解と Holt-Winters 法を組み合わせた CP+HW の 5 つの手法を用いて，提案手法の性能を比較した．また，Truncated SVD，CP 分解のランクおよび NMF の基底 K については，$\mathrm{K}=5,10,\ldots,50$ の範囲で以下の式に準ずる総合値として計算した．

$$\mathbf{S}=\sum_{K\in(5,10,\ldots,50)}\frac{\mathbf{S_K}}{\|\,\mathbf{S_K}\,\|_F} \tag{9}$$

実験では，リンクの有無を予測できたかどうかを識別性能評価に用いられる ROC（Receiver Operating Characteristic）曲線の AUC（Area Under the Curve）によって評価した．

5.2 実験結果

表 1 に ROC 曲線の AUC の値と予測できたリンクの数および精度について各手法について比較した結果を示す．各手法によって，リンクがあると予測された数が予測リンク数であり，その中で実際に正解データの中に存在した割合が適合率である．実際の正解データにあるリンクのうち，どれだけ予測できたかを表すのが再現率である．隠れリンク[*1]とは，予測したリンクの中で過去の観測データに含まれず，正解データにのみ含まれるリンクの割合である．正解データの中の全リンク数は 33501，隠れリンク数は 1260 であった．

表 1　各手法の評価
Table 1　Evaluation result of 6 methods.

評価指標	提案手法	HW	T-SVD CT	T-SVD CWT	CP	CP+HW
AUC	**0.9141**	0.8944	0.8695	0.8915	0.9070	0.9110
全リンク適合率	23.64%	**33.12%**	18.61%	21.31%	23.57%	24.94%
(全リンク正解数)	(28706)	(28275)	(27893)	(28078)	(28504)	(**28783**)
全リンク再現率	85.69%	84.40%	83.26%	83.81%	85.09%	**85.92%**
予測全リンク数	121430	85374	149863	131780	120925	120218
隠れリンク適合率	2.06%	0.0%	1.37%	1.65%	2.04%	**2.11%**
(隠れリンク正解数)	(**741**)	(0)	(660)	(688)	(617)	(646)
隠れリンク再現率	**58.86%**	0.0%	52.38%	54.6%	49.0%	51.29%
予測隠れリンク数	35959	0	48303	41758	30322	30649

表 1 を見ると，提案手法が最も AUC が大きく，識別性能が高いことが分かる．このことから，NMF の特徴抽象化と Holt-Winters 法の周期性予測がうまく機能したと考えられ，複数利用者の行動の共通する特徴の抽出と，これらの特徴に関する周期性のモデル化ができているといえる．一方，全リンク適合率は HW が最も高い結果になった．これは，Holt-Winters 法を用いて単純に各要素を予測する手法だと，過去データに現れるデータのみを予測するので，予測リンク数が少なくなるため適合率が上がったと考えられる．その反面，HW 手法では隠れリンクをまったく予測することができない．その点では，提案手法が正解データにのみ含まれる隠れリンクの再現率が最も高かった．隠れリンクの適合率では，CP+HW 手法が最も高い性能であったが，提案手法と僅差であった．この点で，提案手法は隠れリンクを多く予測する性能が高いといえる．これは，NMF の特徴抽象化によって，似た嗜好を持つ他のユーザの行動から，ユーザの隠れた嗜好を予測できた結果だと考えられる．したがって，グラフ構造の解析による隠れ因子の発見という点において，提案手法が高い性能を持つといえる．

6. 結論

本論文では，時系列リンク予測問題について取り組み，NMF と Holt-Winters 法を組み合わせた手法を提案した．評価実験の結果，時系列リンク予測問題において，提案手法が高い予測性能を持ち，特に潜在的なリンクの予測性能

[*1] 隠れリンクは，単一ユーザの過去の行動から予測することができないので，予測することが難しいため，評価指標として使われている [2].

© 2017 Information Processing Society of Japan

が高いことが確認された．

提案手法が持つ課題あるいは拡張の余地として，NMF の基底数 K や Holt-Winters の予測周期数 m の決定の自動化，NMF の初期値の改善などがあげられる．そのために，K と m を同時に最適化するコスト関数の検討や NMF に対して様々な初期値を与えて実験を行う必要がある．

参考文献

[1] Chatfield, C. and Yar, M.: Holt-winters forecasting: Some practical issues, *J. Royal Statist. Soc. Series D* (*The Statistician*), Vol.37, No.2, pp.129–140 (1988).

[2] Dunlavy, D.M., Kolda, T.G. and Acar, E.: Temporal Link Prediction Using Matrix and Tensor Factorizations, *ACM Trans. Knowledge Discovery from Data*, Vol.5, No.2 (2011).

[3] Liben-Nowell, D. and Kleinberg, J.: The link-prediction problem for social networks, *Journal of the American Society for Information Science and Technology*, Vol.58, No.7 (2007).

[4] Dumais, S.T., Furnas, G.W., Landauer, T.K., Deerwester, S. and Harshman, R.: Using latent semantic analysis to improve access to textual information, *Proc. SIGCHI* (1988).

[5] Getoor, L. and Diehl, C.P.: Link mining: A survey, *ACM SIGKDD Explor. Newslett*, Vol.7, No.2 (2005).

[6] Huang, Z., Li, X. and Chen, H.: Link prediction approach to collaborative filtering, *Proc. JCDL* (2005).

[7] Kalekar, P.S.: *Time series forecasting using Holt-Winters exponential smoothing*, Kanwal Rekhi School of Information Technology (2004).

[8] 亀岡弘和：非負値行列因子分解，計測と制御，Vol.51, No.9 (2012).

[9] 鹿島久嗣：ネットワーク構造予測，人工知能学会誌，Vol.22, No.3 (2007).

[10] Kolda, T.G. and Bader, B.W.: Tensor decompositions and applications, *SIAM Rev.*, Vol.51, No.3 (2009).

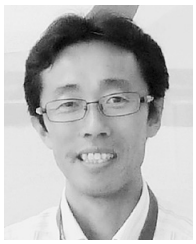
[11] 澤田　宏：非負値行列因子分解 NMF の基礎とデータ/信号解析への応用，電子情報通信学会誌，Vol.95, No.9 (2012).

中嶋 篤宏

2016 年大阪大学工学部電子情報工学科卒業．同大学大学院情報科学研究科在学．

佐々木 勇和

大阪大学大学院情報科学研究科助教．2014 年大阪大学情報科学研究科博士後期課程修了．情報科学博士．データベースシステム，モバイルネットワーク，データマイニングに関する研究に従事．



鬼塚 真

大阪大学大学院情報科学研究科教授．1991 年東京工業大学情報工学科卒業．同年，NTT 入社．2000–2001 年ワシントン大学客員研究員，2013–2014 年電気通信大学客員教授，2012–2014 年 NTT 特別研究員等を経て現職に至る．博士（工学）．2004 年情報処理学会山下記念賞．2008 年日本データベース学会上林奨励賞．2013 年電子情報通信学会論文賞，情報処理学会論文賞．2014 年電子情報通信学会 I-Scover チャレンジ最優秀賞，日本データベース学会論文賞．2015 年電子情報通信学会論文賞．2017 年日本データベース学会論文賞．ACM，電子情報通信学会，情報処理学会各会員．

（担当編集委員　清田 陽司）

© 2017 Information Processing Society of Japan