

# 非負値行列分解を用いた時系列変化に基づくリンク予測

中嶋 篤宏<sup>†</sup> 竹内 孝<sup>††</sup> 佐々木勇和<sup>†</sup> 鬼塚 真<sup>†</sup>

<sup>†</sup> 大阪大学大学院情報科学研究科

<sup>††</sup> NTT コミュニケーション科学基礎研究所上田特別研究室

E-mail: <sup>†</sup>{nakashima.atsuhiko,sasaki,onizuka}@ist.osaka-u.ac.jp, <sup>††</sup>takeuchi.koh@lab.ntt.co.jp

あらまし 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した。さらに、ユーザの基底ベクトルを時間方向に固定する拡張を行うことで、予測精度を向上させることが可能であることが確認された。

キーワード リンク予測, 非負値行列分解, Holt-Winters 予測

## 1. はじめに

近年、情報処理技術の発達により実世界で扱われる膨大なデータを処理することが可能になっている。それらの膨大なデータの中から有益な情報を発見することで、既存のサービスやアプリケーションの性能を高めたり新たな利益を生み出すことが出来る。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。例えば、二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、Facebook などのソーシャルネットワークサービスでは新しい友人を推薦したり、Amazon などの E-commerce サイトであればユーザの嗜好に合った商品を推薦し売り上げを増加させたりすることができる [6], [10], [13], [14]。ソーシャルネットワークや商品の販売乗法などの実世界のデータは、時間とともに変化するため、時間的な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], [21]。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。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 [4], [9]。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捉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 [20]。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 [20]。一方、データの周期的な変化を捉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捉えた予測が可能になる。

本論文では非負値行列分解と Holt-Winters 法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捉えることで、行動を予測する手法を提案する。提案手法では、各時刻のグラフ構造を表す隣接行列を非負値行列分解を用いて 2 つの行列に分解し、分解された行列の各要素に対して Holt-Winters 法を用いて予測することで将来の分解行列を求め、その行列積をとることで将来のグラフ構造を予測する。その際、分解行列の基底ベクトルの順序が時刻間で変化しないように、各時刻の隣接行列に対して時間方向に平均をとった行列を初期行列として NMF を行うことで、基底ベクトルが全時刻に渡って同じ順序で出現するように誘導した。また、時系列グラフをより忠実にモデル化するために、分解行列の基底ベクトルを時間方向に固定する拡張手法を提案する。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 F 値において従来手法を約 1% 上回り、潜在リンクにおける F 値では約 0.2% 上回ることを確認した。また、提案手法と拡張手法の性能を比較した結果、拡張を行うことで提案手法の性能の向上が期待できることを確認した。

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。時系列リンク予測問題のアプローチとして、グラフ構造の特徴を解析する行列分解手法を用いたアプローチと時系列の周期性をとらえる予測手法を用いたアプローチが用いられる。

### 2.1 行列分解を用いたアプローチ

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、二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 [3]。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。

$$\mathbf{X} \approx \mathbf{U}_K \Sigma_K \mathbf{V}_K^T \quad (1)$$

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 [4]。CT は、3 次元テンソルの時系列情報をなくし、時間方向の総和をとる方法である。

$$\mathbf{X} = \sum_{t=1}^T \mathbf{Z}_t \quad (2)$$

CWT は時間的な重みを与えることで、3 次元テンソルの時系列情報を壊さずに 2 次元行列に次元削減する。

$$\mathbf{X} = \sum_{t=1}^T (1 - \theta)^{T-t} \mathbf{Z}_t \quad (3)$$

この次元削減された行列に対して T-SVD を適用することで時系列リンク予測を行う。

また、テンソルベースの時系列リンク予測手法として、Canonical Polyadic(CP) 分解がある [4], [8]。CP 分解は、1 つの 3 次元テンソルを 3 つのベクトルの外積を用いて近似する。

$$\mathcal{Z} \approx \sum_{k=1}^K \lambda_k \mathbf{a}_k \circ \mathbf{b}_k \circ \mathbf{c}_k \quad (4)$$

ここで、記号  $\circ$  は外積を表す<sup>(注1)</sup>。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。

また、グラフ構造に属性情報を付加した行列を解析する手法に Factorization Machines(FM) がある [16]。FM では属性情報を付加したグラフ構造を特殊な行列で保持し、機械学習によって予測値を求めるパラメータを学習する。属性情報として時系列情報を付加することで、時系列リンク予測に応用することが可能である。

### 2.2 予測手法を用いたアプローチ

時系列データの変化を予測するために様々な手法が研究されている [15]。あるノード間のリンクの時間変化を予測することは単変量の時系列予測問題であり、それを全ノードの組み合わせに対して解くことで将来のリンクを予測することが可能である。

単純な単変量の時系列予測手法として指数平滑法がある [2]。1 次の指数平滑法では、観測データ  $y_1, \dots, y_t$  に対して、以下の式によって最新のデータをより重視した予測を行う。

$$P_{t+1} = \alpha y_t + (1 - \alpha) P_t \quad (5)$$

2 次の指数平滑法では、漸増・漸増傾向をとらえるためのトレンド項  $b_t$  を追加した、以下の式で予測を行う。

$$P_{t+1} = \alpha y_t + (1 - \alpha)(P_t + b_t) \quad (6)$$

$$b_t = \beta(P_t - P_{t-1}) + (1 - \beta)b_{t-1} \quad (7)$$

また、過去のデータから将来のデータを予測する Auto Regressive(AR) モデルも広く用いられている [18]。Vector Auto Regressive(VAR) モデルは、時刻のラグセット  $L$ 、重み  $A_l$ 、およびガウシアンノイズ  $\eta_t$  を用いて、多変量時系列データを予測する。

$$\mathbf{y}_t = \sum_{l \in L} \mathbf{A}_l \mathbf{y}_{t-l} + \boldsymbol{\epsilon}_t \quad (8)$$

このほかにも AR モデルは様々な研究分野で用いられており、社会経済学の分野では空間的な自己相関を回帰フレームワークに組み込んだ Simlutaneous Auto Regressive(SAR) モデルが一般的な手法となっている [12]。コンピュータビジョンの分野では、2 次元の画像において空間自己相関を解析する 2D-AR モデルが提案されている [17]。また、AR モデルをテンソル分解と組み合わせて社会センサデータの時空間予測問題などにも応用されている [11]。

## 3. 事前知識

まず、3.1. 章にて非負値行列分解について述べ、3.2. 章にて 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二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、

(注1): 3 次元における外積  $\mathcal{X} = \mathbf{a} \circ \mathbf{b} \circ \mathbf{c}$  は  $\mathcal{X}(i, j, k) = \mathbf{a}(i)\mathbf{b}(j)\mathbf{c}(k)$  を表す

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す、元の行列内の幾つかの頻出パターンを抽出することが出来る。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  $X$  を  $I \times K$  サイズの非負値行列  $U$  と、 $K \times J$  サイズの非負値行列  $V$  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 [19]。乗法的更新では、ランダムな非負値で初期化した行列  $U, V$  に対して、元の行列  $X$  と分解行列の積  $UV$  の乖離度を計算し、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  $U, V$  を求める。本実験では、データセットとしてヘアサロンチェーンの販売情報を用いており、このようなデータセットはポアソン過程に関連して発生すると考えられるため、ポアソン分布に属するデータに適した KL divergence 乖離度規準を用いる。

$$u_{ik} \leftarrow u_{ik} \frac{\sum_j x_{ij} v_{kj}}{\sum_k u_{ik} v_{kj}} \quad (9)$$

$$v_{kj} \leftarrow v_{kj} \frac{\sum_i x_{ij} u_{ik}}{\sum_k u_{ik} v_{kj}} \quad (10)$$

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みなすことができる。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。

### 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 法には加算的方法と乗算的方法があり、本実験で用いたデータセットは季節変動が全観測データを通して一定であると考えられるため、加算的 Holt-Winters 法を用いる。加算的 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。

$$y_{t+h} = l_t + hb_t + s_{t-m+h}. \quad (11)$$

$$l_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(l_{t-1} + b_{t-1}). \quad (12)$$

$$b_t = \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1}. \quad (13)$$

$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1 - \gamma)s_{t-m}. \quad (14)$$

$y_1, y_2, \dots, y_t$  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、 $\alpha, \beta, \gamma$  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、時間毎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捉え、未来の行動を予測する。

### 4.1 基本的な提案手法

図 1 に基本的な提案手法の概要図を示す。

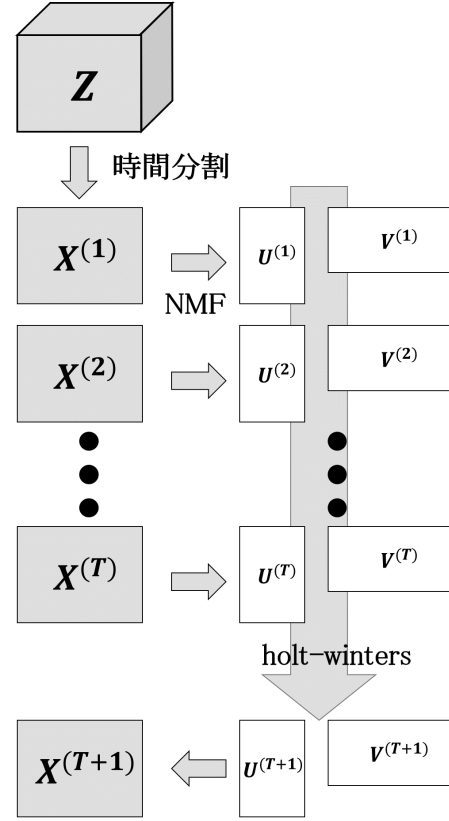


図 1: 提案手法概要

提案手法ではまず、グラフ構造の時間変化を表す 3 次元テンソル  $Z$  に対して、時間軸でデータを均等な時間幅で  $T$  個に分割し、得られた各行列  $X^{(1)}, \dots, X^{(T)}$  に対して NMF を適用することで、分解行列  $U^{(1)}, V^{(1)}, \dots, U^{(T)}, V^{(T)}$  を得る。そして、時間方向に連続する各  $U^{(t)}, V^{(t)}$  の各要素に対して Holt-Winters 法を適用し、 $U^{(T+1)}, V^{(T+1)}$  を予測し、 $X^{(T+1)} = U^{(T+1)}V^{(T+1)}$  によって最終的な予測行列  $X^{(T+1)}$  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 ( $U$  ならば行ベクトル、 $V$  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行列  $X_{Ave}(i, j) = \frac{1}{T} \sum_{t=1}^T X^{(t)}(i, j)$  に対して、NMF を適用して得られた分解行列を、全時間軸にわたって共通の初期分解行列  $U_{init}, V_{init}$  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誘導することができる。これにより、時間軸全体の行列の特徴を捉えることが期待できる。

Algorithm1 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  $X$  と分割されたデータの時間方向の平均行列  $X_{Ave}$  を入力として、1 単位時間先の予測行列  $X^{(T+1)}$  を出力する。平均行列にランダムな初期値  $U_{random}, V_{random}$  を用いて NMF を適用し、初期分解行列  $U_{init}, V_{init}$  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  $U^{(T+1)}, V^{(T+1)}$  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  $S = X^{(T+1)}$  を求める (9 行目)。

---

#### Algorithm 1 予測行列 $X^{(T+1)}$ の計算

---

Input  $X = (X^{(1)}, X^{(2)}, \dots, X^{(T)}), X_{Ave}$

Output  $X^{(T+1)}$

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

#### 4.2 要素を固定した行列分解手法

2 部グラフにおける時系列リンク予測では、2 つの部分集合のうち 1 つは時間的に不変で、もう 1 つが時間的な変化を持つといったことが考えられる。例えば、ユーザの嗜好は不変であるが、商品の特性が時間的に変化する場合、あるいはその逆の場合などが考えられる。そこで、提案手法の行列分解を改良して、行列の列要素を固定し行方向に行列  $X^1, X^2, \dots, X^{T+1}$  を連結させた行列に対して行列分解を行う図 2 のような手法を考える。

この手法ではまず、行方向に時間順に  $I \times J$  サイズの行列を  $T$  個連結させた  $TI \times J$  サイズの行列に対して非負値行列分解を行い、 $TI \times K$  サイズ行列  $U$  と  $K \times J$  サイズ行列  $V$  の 2 つの行列に分解する。こうすることで、分解後の列要素は全時間に渡って同じ値を持ち、行要素のみによって時間変化が表される。また、全時間に渡って列要素が同じため、列要素の特徴因子の出現場所が変わるといった問題が起こらないという利点がある。そして、 $TI \times K$  行列  $U$  を行方向に  $T$  個に分割し、 $U^1, U^2, \dots, U^T$  とし、基本的な提案手法と同様に  $U^1, \dots, U^T$  に対して Holt-Winters 予測を行うことで、 $U^{T+1}$  を得る。この行要素の時間変化の予測行列  $U^{T+1}$  と、全時間に渡って同一の  $V$  の積をとることで、最終的な予測行列  $X^{T+1}$  を得る。

Algorithm2 に列要素を固定した提案手法の全体的なアルゴリズムを示す。行方向に時間順に連結された行列  $X$  を入力とし

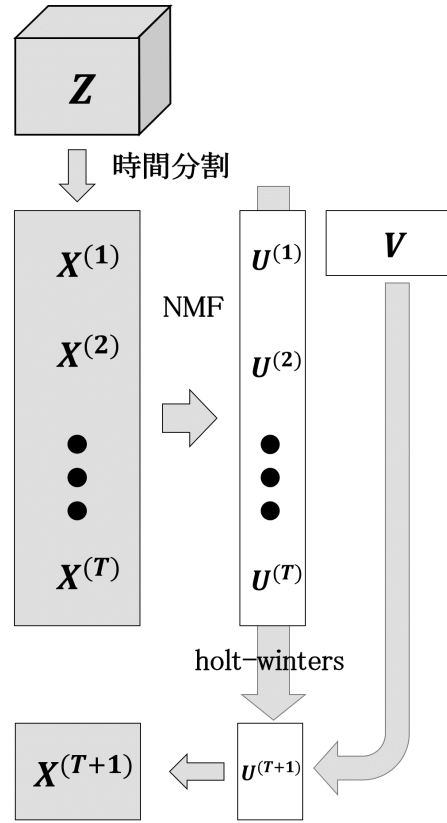


図 2: 列要素を固定した手法概要

て、1 単位時間先の予測行列  $X^{(T+1)}$  を出力する。 $TI \times J$  サイズの行列  $X$  に対して非負値行列分解を行い、 $TI \times K$  サイズ行列  $U$  と  $K \times J$  サイズ行列  $V$  の 2 つの行列に分解する (1 行目)。 $TI \times K$  行列  $U$  を行方向に  $T$  個に分割し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  $U^{(T+1)}$  を求める (2 行目から 4 行目)。予測した分解行列  $U^{T+1}$  と全時間に渡って同一の分解行列  $V$  の積を計算し、1 単位時間先の予測行列  $S = X^{(T+1)}$  を求める (5 行目)。

---

#### Algorithm 2 列要素を固定した予測行列 $X^{(T+1)}$ の計算

---

Input  $X = (X^{(1)}, X^{(2)}, \dots, X^{(T)})$

Output  $X^{(T+1)}$

- 1: NMF :  $U, V \leftarrow X$
  - 2: Time Divide :  $U^{(1)}, U^{(2)}, \dots, U^{(T)} \leftarrow U$
  - 3:  $Ulist \leftarrow U^{(1)}, U^{(2)}, \dots, U^{(T)}$
  - 4: Holt - Winters :  $U^{(T+1)} \leftarrow Ulist$
  - 5:  $X^{(T+1)} = U^{(T+1)}V$
- 

行要素を固定したい場合は、列方向に時間順に連結させた行列を行列分解することで同様にを行うことができる。2 部グラフの一方のノードのみに時間的な変化要因が含まれている、もしくは一方の変化要因の影響が他方に比べて著しく大きい場合は、この手法によりデータをより忠実なモデル化し予測精度が高くなることが期待できる。

## 5. 評価実験

この章では、4. 章にて述べた提案手法と 2. 章にて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。また、4. 章にて述べた行列要素を固定した手法の優位性について検証した実験と結果について述べる。

### 5.1 既存手法との比較実験

2. 章にて述べた既存手法と 4. 章にて述べた提案手法のリンク予測の精度について比較する。使用したデータセットは、経営科学系研究部会連合協議会主催平成 29 年度データ解析コンペティションで提供された、首都圏のあるヘアサロンチェーンの 24 か月分の販売データである。31862 ユーザと 1399 商品の中から、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、最後の一个月を除く 23 か月のデータを月毎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 [4] 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、

$$Z(i, j, t) = \begin{cases} 1 + \log(n) & n > 0 \\ 0 & n = 0 \end{cases} \quad (15)$$

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、

$$Y(i, j) = \begin{cases} 1 & n > 0 \\ 0 & n = 0 \end{cases} \quad (16)$$

で表される行列  $Y$  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 か月とした。これは、販売情報のようなデータには 4 半期毎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、 $X_1, \dots, X_T$  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、 $K=5, 10, \dots, 50$  の範囲で以下の式に準ずる総合値として計算した。

$$S = \sum_{K \in (5, 10, \dots, 50)} \frac{S_K}{\|S_K\|_F} \quad (17)$$

実験では、リンクの有無を決定する閾値を識別性能評価に用いられる ROC (Receiver Operating Characteristic) 曲線を用いて、最も真陽性率が高く偽陽性率が低い値を選択した。ROC 曲線下の面積 AUC (Area Under the Curve) と F 値によって予測性能を評価した。

図 3, 4 に実験結果を示す。

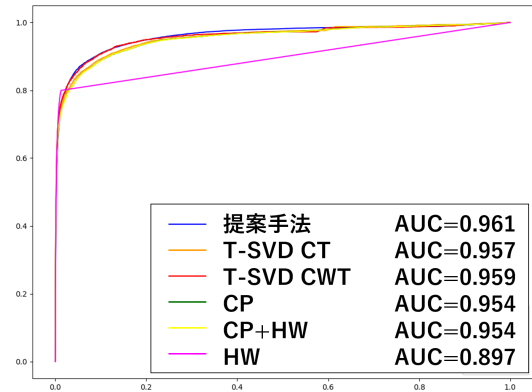
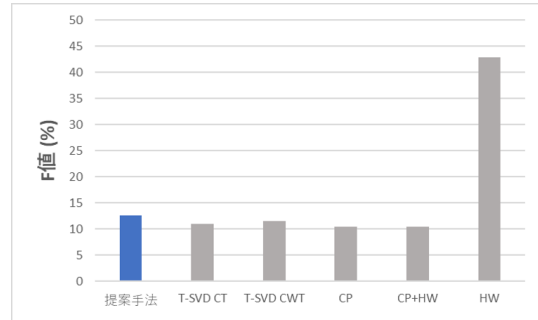
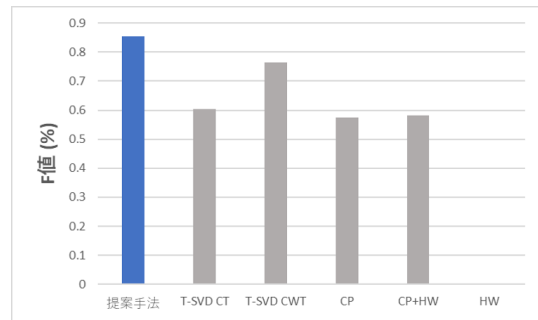


図 3: 各手法の ROC 曲線



(a) 総リンクにおける F 値



(b) 隠れリンクにおける F 値

図 4: 各手法の F 値

隠れリンク<sup>(注2)</sup>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。正解データの中の総リンク数は 33501、隠れリンク数は 1260 であった。

ROC 曲線は左上に近づくほど識別性能が高く、その曲線下の面積 AUC も高くなる。図 3 をみると、HW 手法を除いた手法の ROC 曲線は似た形をして重なっているが、AUC をみると提案手法が最も性能が高いことが分かる。図 4a をみると、HW 手法が最も性能が高かった。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り適合率が上がったためと考えられる。その反面、図 4b をみると、HW 手法では隠れリンクを全く予測することが出来ないため、F 値

(注2): 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 [4]。

は 0 である．提案手法は HW 手法を除いた比較手法の中で最も高い F 値を示している．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ているといえる．また図 4b においても，提案手法が最も高い F 値を示している．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，提案手法は高い性能をもつといえる．

### 5.2 行列要素を固定した手法との比較

4. 章にて述べた行列要素を固定した手法が基本的な提案手法に比べてリンク予測の性能が変化するか検証した．また，行要素（ユーザ成分）と列要素（商品成分）のどちらを固定するほうが性能が高くなるかも同時に検証した．実験条件は，5.1. 節と同様であり，基本的な提案手法と比較とした．

図 5, 6 に実験結果を示す

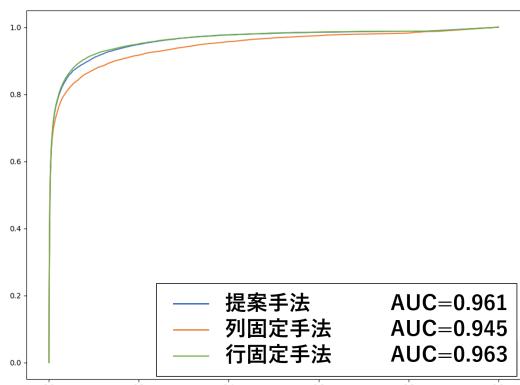
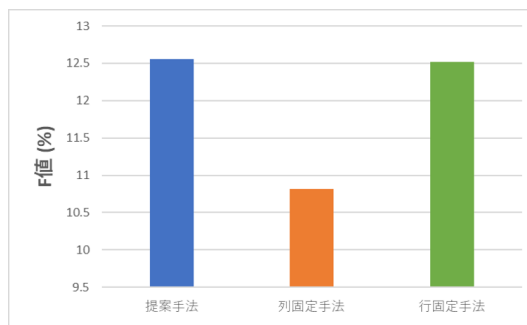
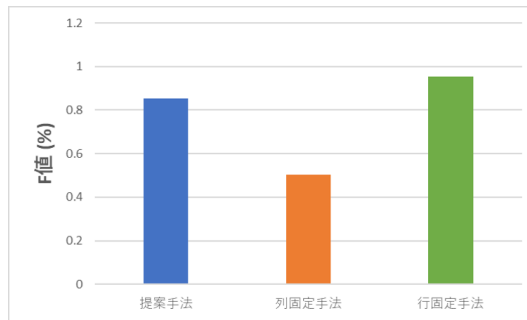


図 5: 要素固定手法の ROC 曲線

図 5 をみると，行固定手法が基本的な提案手法に比べ僅かに左上に近く，その 2 つの手法から列固定手法右下方向に乖離していた．AUC をみても行固定手法が僅差で基本的な提案手法より高く，最も識別性能が高いといえる．一方，図 6a をみると，行固定手法の F 値は基本的な提案手法に比べて僅かに低く，列固定手法は 2 つに比べて大きく低くなっていた．また，図 6b をみると，行固定手法は基本的な提案手法を上回っていた．本データセットにおいて，行要素はユーザを表し，列要素は商品を表す．列固定手法に比べ行固定手法のほうが性能が高いということは，本データセットにおいてはユーザの嗜好は不変であり，商品の特性によって売上が時間的に変化するという性質を持っているため，行要素を固定した手法がより効果的であり，高い性能を示したと考えられる．行固定手法は総リンクの F 値においては，基本的な提案手法に劣るが，AUC および隠れリンクの F 値においては上回っており，データセットによっては基本的な提案手法の性能を上回ることが期待できる．今回使用したデータセットは 1 つであったため，データセットと各手法の相性の関係性は検証できていないが，適用するデータセットの特性を分析したうえで，適切な手法を選択することでより予測性能を高めることができるだろう．



(a) 総リンクにおける F 値



(b) 隠れリンクにおける F 値

図 6: 要素固定手法の性能比較

## 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が高いことが確認された．また，提案手法に行列要素を固定するという拡張を行うことで，予測精度を高めることが期待できる．

本研究の実験では，データセットのサイズが小さく売上数の多い商品の影響や恒常的に購入され周期性をあまり持たない商品の影響が大きく反映されてしまうため，データセットのサイズを大きくしてさらに実験を行うことが必要である．また，他のデータセットにおいても実験を行い，提案手法および行列要素を固定する手法の性能を検証する必要がある．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．

### 文 献

- [1] Chris Chatfield and Mohammad Yar. Holt-winters forecasting: some practical issues. *J. Royal Statist. Soc. Series D (The Statistician)*, Vol. 37, No. 2, pp. 129–140, 1988.
- [2] Frank Chen, Jennifer K. Ryan, and David Simchi-Levi.
- [3] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. Using latent semantic analysis to improve access to textual information. *In Proceedings of SIGCHI*, 1988.
- [4] Daniel M. Dunlavy, Tamara G. Kolda, and Evrim Acar. Temporal link prediction using matrix and tensor factorizations. *ACM Transactions on Knowledge Discovery from*

*Data*, Vol. 5, No. 2, 2011.

- [5] L. Getoor and C. P. Diehl. Link mining: A survey. *ACM SIGKDD Explor. Newslett*, Vol. 7, No. 2, 2005.
- [6] Zan Huang, Xin Li, and Hsinchun Chen. Link prediction approach to collaborative filtering. In *Proceedings of JCDL*, 2005.
- [7] P. S. Kalekar. Time series forecasting using holt-winters exponential smoothing. *Kanwal Rekhi School of Information Technology*, 2004.
- [8] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Rev*, Vol. 51, No. 3, 2009.
- [9] Yahuda Koren. Collaborative filtering with temporal dynamics. *KDD '09 Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 447–456, 2009.
- [10] Yahuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, Vol. 42, No. 8, pp. 30–37, 2009.
- [11] K. Takeuchi, H. Kashima, and N. Ueda. Autoregressive tensor factorization for spatio-temporal predictions. *IEEE International Conference on Data Mining (ICDM)*, 2017.
- [12] L. Anselin and A. K. Bera. Spatial dependence in linear regression models with an introduction to spatial econometrics. *Statistics Textbooks and Monographs*, Vol. 155, pp. 237–290, 1986.
- [13] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 7, 2007.
- [14] Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, Vol. 7, , 2003.
- [15] Spyros Makridakis and Michele Hibon. The m3-competition: results, conclusions and implications. *International Journal of Forecasting*, Vol. 16, pp. 451–476, 2000.
- [16] Steffen Rendle. Factorization machines. *ICDM '10 Proceedings of the 2010 IEEE International Conference on Data Mining*, pp. 995–1000, 2010.
- [17] S. R. Dubois and F. H. Glanz. An autoregressive model approach to two-dimensional shape classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, No. 1, pp. 237–290, 1986.
- [18] Hsiang-Fu Yu, Nikhil Rao, and Inderjit S. Dhillon.
- [19] 亀岡弘和. 非負値行列因子分解. 計測と制御, No. 9, 2012.
- [20] 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.
- [21] 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.