DEIM2022 D23-3

Joint Non-negative Tensor Factorization に基づく 共通・非共通トピックとその時間推移の抽出

平澤　嶺† 伊藤　寛祥†† 松原　正樹†† 森嶋　厚行††

† 筑波大学情報学群情報メディア創成学類　〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学　〒 305–8550 茨城県つくば市春日 1-2
E-mail: †rei.hirasawa.2021b@mlab.info, ††{ito,masaki,mori}@slis.tsukuba.ac.jp

あらまし　文書データの増大に伴い，大量の文書データの内容を効率的に把握することは重要な課題となっている．例えば，日々大量の特許が出願・取得されているが，これらの特許に目を通して技術動向を把握することは大量であるが故に困難である．これらの文書は，多くの場合，時間情報を持ち，複数のサブ集合 (所属ドメインや著者情報などによる分類) に分けられる．したがって，時間情報を考慮してサブ集合間のトピックを比較することは，大規模文書データを詳細に理解するために有効な手段の一つだと考えられる．本研究では二つの文書集合間で共通するトピックと非共通のトピック，それらのトピックの時間推移情報を抽出する手法を提案する．本手法は，時系列文書データを単語-文書-時間の 3 階の非負値テンソルとして扱い，非負値テンソル因子分解 (Non-negative Tensor Factorization) を応用して二つの集合間で比較しながら分解することで，上述のようなトピックと時間推移の抽出を実現する．本稿では，実験として正解が自明な生成データと建築分野の特許文書データに本手法を適用した結果を示す．

キーワード　データマイニング，トピック抽出，時系列文書

1 はじめに

流通する文書データの量が増大し，複雑な大規模時系列文書データを効率的に把握して全体像を理解する重要性が増している．たとえば AP 通信はニュースの消費者へのインタビューによって，昨今の大量の情報は視聴者が多くの洞察を得ることの役に立っていないことを発見した [9]．ここでインタビューの参加者はニュースが少ないほうが必要な情報を見つけることができると説明している．

ニュースや特許，論文などの時間情報をもつ大規模文書の集合は多くの場合，発行元のドメインや著者の性別・読者の年齢層や国籍などといった情報によって，サブ集合に分けることができ，これらのサブ集合について時間による変化を考慮しながら比較しなければ大規模時系列文書データの把握が難しい場合がある．より具体的には，大量の特許データを分析する際に企業 a が出願した特許の文書集合と企業 b が出願した特許の文書集合にそれぞれどのような技術トピックが含まれていて，その技術の推移が企業間でどう相違しているか，などと言った洞察を得たい場合が挙げられる．また，各国での COVID-19 に関する報道内容の時間推移の比較やネット記事における異なる年齢層に人気の話題と時間推移の比較などにもニーズがあると考えられる．

本論文では (1) 文書内容の分析 (2) データ集合間の比較 (3) 時間変化の抽出，の三つを満たすトピック抽出の手法を提案する．大規模文書データからトピックを抽出して内容を把握する手法として確率モデルに基づく Blei らによる Latent Dirichlet Allocation(LDA) [2] がよく知られており，時系列データに対応した Dynamic topic models [1] など多くの拡張手法につながっている．しかし，複数の文書集合間で内容を比較して，特徴の類似や相違を知りたいといったニーズに対してこれらの手法は十分ではない．また，文書集合の比較をすることができる手法としては Joint Non-negative Matrix Factorization(Joint-NMF) [5] があるが，こちらは時間情報を考慮した分析に対応していない．このように，(1) 文書内容の分析 (2) データ集合間の比較 (3) 時間変化の抽出，の三つを満たす分析は今ある既存手法では難しい．

以上のようなニーズと既存手法の不足を背景として本研究では，Non-negative Tensor Factorization(NTF) [10] と Joint Non-negative Matrix Factorization(Joint-NMF) [5] の二つの手法を組み合わせることで，二つの時系列文書集合間で共通する話題と各文書集合に固有の話題とそれらの時間推移の同時抽出を実現する．NTF とは非負値行列におけるトピック抽出の手法である Non-negative Matrix Factorization(NMF) [6] を高階の非負値テンソルを扱えるように拡張した手法で画像解析 [4] や関係データの分析に用いられている．本論文ではこの NTF の手法を時系列文書に当てはめるために，時系列文書データをテンソルとして扱う新しいデータ構造を提案する．Joint-NMF は非負値行列として表現された二つのデータを比較しながら行列分解できるように NMF の損失関数に項を追加したものだが，本論文ではこの手法を参考に非負値テンソルで表現された二つの時系列文書データを時間推移を考慮して比較しながら分解できるような損失関数を提案する．

本論文で提案する手法では，非負値テンソルとして表現された二つの時系列文書集合を入力とし，損失関数を最適化してテンソルを行列の積に分解することで，出力としてそれぞれの文

表 1: 既存手法と提案手法の比較

	文書内容分析	集合間比較	時間変化抽出
DTM	○	-	○
time-collective-NMF	○	-	○
Joint-NMF	○	○	-
NTF	-	-	○
提案手法 (Joint-NMF+NTF)	○	○	○

書集合で共通している複数のトピックと共通していない複数のトピックの単語分布を示す行列，それらのトピックと各文書との関連度合いを示す行列，それらのトピックの時間経過での増減を示す行列を得る．

本論文では正解が自明な生成データと実データ (建築分野の特許文書データ) に本手法を適用した実験の結果を示す．生成データによる実験では，正解データとして共通・非共通成分を持つ行列を生成し，これらの積である X_i を提案手法により再び行列に分解することで，提案手法が正しく共通・非共通成分を抽出できるかを調べた．実データによる実験では，二つの企業の特許文書集合に本手法を当てはめることで，それぞれの企業の特許文書に含まれる共通・非共通の技術トピックとそれらの時期ごとの増減を抽出して結果を調べた．

本論文の構成は以下である．2 章では関連研究について説明を行い，3 章では提案手法の詳細な説明をする．まず時系列文書をテンソルとして扱うための新しいデータ構造について説明し，次にそれらの分解を行う損失関数の説明，最適化の説明を行う．4 章では正解が自明な生成データによる実験結果と建築分野の特許文書データでの実験結果を示す． 最後に 5 章で結論を述べる．

2 関 連 研 究

文書や時系列データの特徴を抽出する手法は非負値行列を分解する NMF の応用として多くの手法が提案されてきた．また，NMF の応用以外にも確率的な生成モデルを考える手法など様々なアプローチがある．しかし，表 1(本研究と特に関係が深い既存手法と提案手法を比較した表) に示すように既存手法は (1) 文書内容の分析 (2) データ集合間の比較 (3) 時間変化の抽出，の三つを満たした分析には対応していない．本章では関連研究を概説し，提案手法との関係や違いを述べる．

2.1 Non-negative Matrix Factorization(NMF) に関連した研究

Non-negative Matrix Factorization(NMF) [6] とは Lee らにより提案された非負値行列を二つの非負値行列の積に分解する手法である．この分解は行列を特徴ベクトルと重みベクトルの直積の線形結合に分解しているとみなすことができ，結果の解釈がしやすいことから様々な発展手法につながっている．本論文での提案手法も NMF の発展型に含まれる．

NMF に基づく手法の中で複数のデータ集合間の比較ができるものとしてはまず，Gupta らによる Regularized nonnegative shared subspace learning [3] が挙げられる．この手法では二つの入力データを比較しながら行列分解をすることで，共通する特徴を検出することができる．しかし，データ集合間の厳密に等しい成分を抽出する手法であるため柔軟性に欠ける．次に，Kim らによる手法 (joint-NMF) [5] が挙げられる．この手法は二つのデータを比較して共通・非共通成分を抽出する. [3] の手法とは異なり，共通・非共通の度合いを調節するためのハイパーパラメータを持つため，柔軟にデータを比較することができる．しかし，入力データの表現は行列に限定されており，提案手法のように文書データに対して，時間情報を加味した分析を行うことはできない．本手法はこの joint-NMF を時間情報を考慮できるように拡張する．

NMF に基づく手法のうち，時間情報を考慮してデータの特徴を把握する手法には，Vaca らによる手法 (time-collective-NMF) [12] がある．この手法は Collective matrix factorization [11] という教師データを用いてコンテキストに沿った行列分解を行う手法の応用である．time-collective-NMF [12] では，過去の検出結果をコンテキストとして，以前検出されたトピックの内容が現在どのように変化しているかが分かるように行列を分解する手法である．この手法で得られる結果はトピックの内容自体がどのように変化したり出現・消失するかを示すものであり，提案手法は同一のトピックが時間経過の中でどのように増減するかを示すという点が異なる．また time-collective-NMF [12] では複数のデータを比較して共通要素や非共通要素を発見することができない．NMF を 3 階以上の非負値テンソルを入力として分解できるように拡張した Non-negative Tensor Factorization(NTF) [10] も時系列データの分析に用いることができる．この手法は画像データ [4], [10] や関係データなど様々な分析に用いられているが，時系列文書データをテンソルとして扱って NTF を適用するためには工夫が必要である．提案手法では時系列文書をテンソルとして扱うための新しいデータ構造を導入して，時系列文書データを入力とした NTF を行う．本研究の技術的な位置づけとしては，行列しか扱うことのできなかった Joint-NMF の手法に NTF の手法組み合わせることでテンソルを扱えるようにし，さらに提案するデータ構造によって時系列文書データの分析を行えるようにしたものであるといえる．

2.2 その他のトピック抽出手法や時系列データにおけるデータマイニング手法

NMF 以外のトピック抽出の手法には確率的なアプローチをとるものもある．Blei らによる Latent Dirichlet Allocation(LDA) [2] が代表であり，データの潜在的なトピックを確率的に推定する．これは著者情報を取り込むための拡張などもされている [7]．また，Blei らは LDA を時系列データ分析のために拡張した Dynamic topic models(DTM) [1] を提案している．Dynamic topic models はトピックの時間推移での増減を抽出できるが，提案手法はデータ集合間の共通点や非共通点を検出した上でそれらの時間推移を把握することを目標

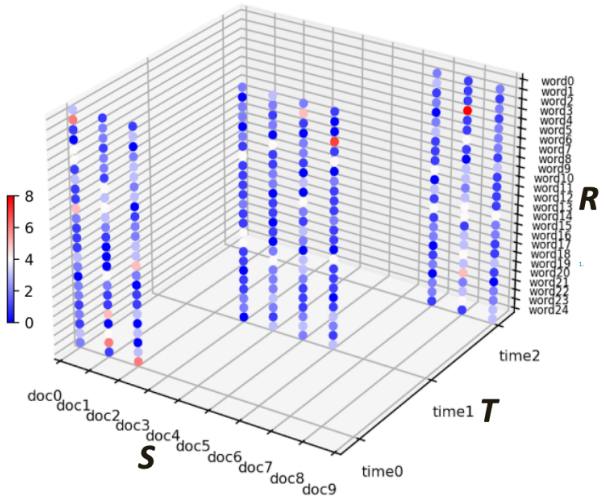


図 1: 本手法で用いる非負値テンソルの構造の例 (R=25，S=10，T=3)

としており，この手法と異なる．テンソル表現を用いているが NMF の枠組みには入らない時系列データの分析の手法としては Matsubara らによる Non-linear mining of competing local activities(CompCube) [8] なども挙げられる．これは検索クエリ数などの時系列データを比較しながら頻出する時系列パターンや外れ値などを検出する手法であるが，文書からのトピック抽出や時間推移の抽出自体は行わない．

3 提 案 手 法

本章では提案手法の詳細を説明する．提案手法では，まず二つの大規模時系列文章集合から 3 階の非負値テンソル $\mathbf{X}_1 \in \mathbb{R}_+^{R\times S_1\times T}$，$\mathbf{X}_2 \in \mathbb{R}_+^{R\times S_2\times T}$ を作る．その後，Non-negative Tensor Factorization を応用した手法で，双方を比較しながら行列に分解する損失関数を最小化する．これにより二つのテンソルを共通・非共通成分を持った行列の積に分解し，二つの文書集合間で共通トピック Kc 個と非共通トピック Kd 個，さらにそれらのトピックの時間推移での増減を抽出する．

3.1 時系列文書集合の非負値テンソル化

本研究で扱うデータは比較したい二つの時系列文書集合である．それぞれを 3 次元非負値テンソル $\mathbf{X}_1 \in \mathbb{R}_+^{R\times S_1\times T}$，$\mathbf{X}_2 \in \mathbb{R}_+^{R\times S_2\times T}$ で表現して，双方を比較しながら共通・非共通トピックとその時間推移を抽出する．それぞれのテンソル X_i は R 個の単語，S 個の文書，T 個の時間区分の 3 軸で構成され，$\mathbf{X}$ の要素 X_{rst} は時間区分 t の文書 s に単語 r がどの程度出現したかを数値化した非負の値で示す．この特徴量としては tf-idf などの値を用いる．

テンソルは例として図 1 のような構造になる．図 1 は R=25，S=10，T=3 の 3 次元テンソルの例である．テンソル上のそれぞれのドットが (時間区分 t，文書 s，単語 r) の特徴量 (tf-idf など) の値を表す．文書というデータの性質上，それぞれの文書は一つの時間区分のみに属するため，その文書が存在しない時間区分の領域については値を持たない．図 1 のように，値を

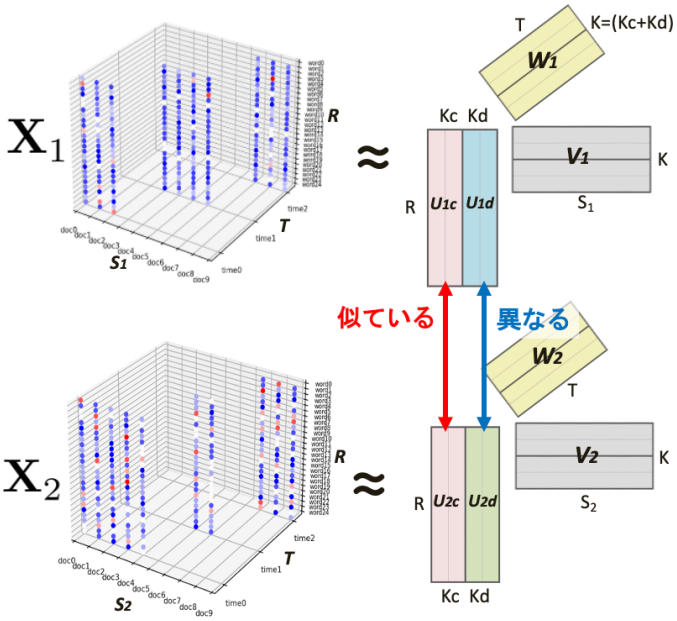


図 2: 本手法におけるテンソルの行列への分解

持たない領域 (ドットがない領域) については要素がないテンソルとなる．このとき，文書 s が属する時間区分 t を $t^{(s)}$ と表し，はじめの時間区分から時間区分 t までに属する文書の数の合計を $S^{(t)}$ と表すことにする．

3.2 Non-negative Tensor Factorization(NTF)

本節では，提案手法のベースとなる Non-negative Tensor Factorization(NTF) について説明する．NTF とは非負値テンソルを低ランクの非負値の行列積で近似する分解計算の手法である．たとえば，3 階の非負値テンソル $\mathbf{X}$ を非負値行列 $U \in \mathbb{R}_+^{R\times K}$，$V \in \mathbb{R}_+^{S\times K}$，$W \in \mathbb{R}_+^{T\times K}$ の積に近似する．つまり $\mathbf{X}$ を

$$\mathbf{X} \approx UVW \tag{1}$$

となるような U, V, W に分解する．ここで三つの行列の積は以下のように定義される．

$$UVW = \sum_{k=1}^{K} U_{(:,\,k)} \otimes V_{(:,\,k)} \otimes W_{(:,\,k)} \tag{2}$$

$\otimes$ はベクトルの直積を意味し，(:, k) は行列の列を k 番目に固定して行方向にインデックスをすべて動かすことで行列から k 列目の列ベクトルを得ることを意味する．このような分解は解析的に解くことができないため，NTF の式 (1) の左辺 $\mathbf{X}$ と右辺 UVW からできる $\mathbf{X}$' の差異を損失関数として最適化により近似する．二乗誤差を用いると損失関数の最適化問題は

$$\underset{U,\,V,\,W\geqq 0}{\arg\min} \sum_{t=1}^{T}\sum_{r=1}^{R}\sum_{s=1}^{S}\left(\mathbf{X}_{(r,\,s,\,t)} - \sum_{k=1}^{K} U_{(r,\,k)}V_{(s,\,k)}W_{(t,\,k)}\right)^2 \tag{3}$$

となる．これにより得られた行列 U, V, W から，入力の $\mathbf{X}$ の特徴を知ることができる．

3.3 提案手法におけるテンソルの分解と損失関数の定義

次に NTF を応用した提案手法を説明する．図 2 に示すよう

に二つの時系列文書集合から作った非負値テンソル $\mathbf{X}_i$ を NTF の手法でそれぞれ三つの非負値行列 $U_i \in \mathbb{R}_+^{R\times K}$，$V_i \in \mathbb{R}_+^{S_i\times K}$，$W_i \in \mathbb{R}_+^{T\times K}$ に分解する．(i は 1 または 2，K はトピック数)．この分解結果の U_i(トピック行列) を見ることで，それぞれのサブ文書集合に含まれていたトピックの単語分布を知ることができる．V_i(文書行列) を見ることで文書集合の各文書がどのトピックにどの程度関連しているかを知ることができる．W_i(時間行列) を見ることでそれぞれのトピックが時間経過の中でどのように増減しているかの推移を知ることができる．これと同時に，トピックを表す行列 (U_i) の一部 (U_ic) については二つの時系列文書集合間でなるべく似ているトピックを表し，残りの部分 (U_id) についてはなるべく異なる特徴を表すように分解する．これにより Kc 個の共通トピックと Kd 個の非共通トピックを抽出する．この分解ための損失関数の最適化問題は NTF の最適化問題である式 (3) の拡張であり，次のようになる．

$$
\begin{aligned}
&\underset{U_1,\,V_1,\,W_1,\,U_2,\,V_2,\,W_2\geqq 0}{\arg\min} L \\
&= \frac{1}{S_1}\sum_{t=1}^{T}\sum_{r=1}^{R}\sum_{s=S_1^{(t-1)}}^{S_1^{(t)}}\left(\mathbf{X}_{1(r,\,s,\,t)} - \sum_k U_{1(r,\,k)}V_{1(s,\,k)}W_{1(t,\,k)}\right)^2 \\
&+ \frac{1}{S_2}\sum_{t=1}^{T}\sum_{r=1}^{R}\sum_{s=S_2^{(t-1)}}^{S_2^{(t)}}\left(\mathbf{X}_{2(r,\,s,\,t)} - \sum_k U_{2(r,\,k)}V_{2(s,\,k)}W_{2(t,\,k)}\right)^2 \\
&+ \alpha\|U_{1c} - U_{2c}\|_F^2 + \beta\|U_{1d}^T U_{2d}\|_1^1 \\
&s.\,t.\ \|U_{1(:,\,k)}\|_2 = \|U_{2(:,\,k)}\|_2 = 1 \\
&\qquad \|V_{1(S_1^{(t-1)}:S_1^{(t)},\,k)}\|_1 = \|V_{2(S_2^{(t-1)}:S_2^{(t)},\,k)}\|_1 = 1 \\
&\quad \text{for } k = 1,\cdots,K \text{ for } t = 1\cdots,T
\end{aligned}
\tag{4}
$$

この式 (4) の第一項と第二項は NTF の式 (3) と基本的に同様であるが，3.1 節で述べたように本手法では文書データを扱っており，一つの文書は一つの時間にしか属してしないため入力のテンソル内に要素を持たない領域がある．そのため入力のテンソルが要素を持つ部分だけについて最適化すべきであり，損失関数のテンソル・行列の添字 s の範囲を指定している．第三項と第四項はそれぞれ，文書データ間での共通トピックと非共通トピックが現れるようにするための項である．第三項はそれぞれの文書集合のトピック行列内の U_{1c} と U_{2c} の差のフロベニウスノルムを示し，この項により U_{1c} と U_{2c} が似た行列になるので文書集合間での共通トピックを表すように分解される．第三項はそれぞれの文書集合のトピック行列内の U_{1d} と U_{2d} について, 転置した U_{1d} と U_{2d} の積の全要素の和を表しており，この項により U_{1d} と U_{2d} はトピックの単語分布がなるべく異なるように最適化されるため非共通トピックが抽出される．α と β は第三項と第四項の式 (4) における影響力を調整することで共通・非共通の厳密さを調節するためのハイパーパラメータである．制約条件はトピックの単語分布を示すベクトルとその文書ごとの重みを示すベクトルをそれぞれ正規化することで，抽出されたトピックとその重みの時間推移について比較可能で解釈しやすいような結果を得るためのものである．

3.4 最適化アルゴリズム

式 (4) は、すべてのパラメータについて同時に凸にはならない。しかし、[6] で述べられている方針に従って，それぞれのパラメータごとに乗法更新を繰り返すことで式 (4) の局所最小値を求めることができる。この乗法更新式は Karush-Kuhn-Tucker(KKT) 条件の第一条件を式 (4) の損失関数 L に適応し，式変形することで得ることができる．行列 $U_{ic}, U_{id}, V_i, V, W_i$ の各要素について，導出された乗法更新式は以下のようになる．

$$
U_{ic(r,k)} = U_{ic(r,k)} \times \frac{\frac{1}{S_1}\sum_{t=1}^{T}\sum_{s=S_i^{(t-1)}}^{S_i^{(t)}}\left(\mathbf{X}_{i(r,\,s,\,t)}V_{ic(s,\,k)}W_{ic(t,\,k)}\right) + \alpha U_{jc(r,k)}}{\frac{1}{S_1}\sum_{t=1}^{T}\sum_{s=S_i^{(t-1)}}^{S_i^{(t)}}\left[V_{ic(s,\,k)}W_{ic(t,\,k)}\sum_{k'=1}^{K}\left(U_{ic(r,\,k')}V_{ic(s,\,k')}W_{ic(t,\,k')}\right)\right] + \alpha U_{ic(r,k)}}
$$

$$
U_{id(r,k)} = U_{id(r,k)} \times \frac{\frac{1}{S_1}\sum_{t=1}^{T}\sum_{s=S_i^{(t)}}^{S_i^{(t)}}\left(\mathbf{X}_{i(r,\,s,\,t)}V_{id(s,\,k)}W_{id(t,\,k)}\right) + \beta U_{jd(r,k)}}{\frac{1}{S_1}\sum_{t=1}^{T}\sum_{s=S_i^{(t)}}^{S_i^{(t)}}\left[V_{id(s,\,k)}W_{id(t,\,k)}\sum_{k'=1}^{K}\left(U_{id(r,\,k')}V_{id(s,\,k')}W_{id(t,\,k')}\right)\right] + \beta\sum_{k'=1}^{K}U_{jd(r,k')}}
$$

$$
V_{i(s,k)} = V_{i(s,k)} \times \frac{\sum_{r=1}^{R}\sum_{t=1}^{T}\left(\mathbf{X}_{i(r,\,s,\,t)}U_{i(r,\,k)}W_{i(t,\,k)}\right)}{\sum_{r=1}^{R}\sum_{t=1}^{T}\left[U_{i(r,\,k)}W_{i(t,\,k)}\sum_{k'=1}^{K}\left(U_{i(r,\,k')}V_{i(s,\,k')}W_{i(t,\,k')}\right)\right]}
$$

$$
W_{i(t,k)} = W_{i(r,k)} \times \frac{\sum_{r=1}^{R}\sum_{s=S_i^{(t-1)}}^{S_i^{(t)}}\left(\mathbf{X}_{i(r,\,s,\,t)}U_{i(r,\,k)}V_{i(s,\,k)}\right)}{\sum_{r=1}^{R}\sum_{s=S_i^{(t-1)}}^{S_i^{(t)}}\left[U_{i(r,\,k)}V_{i(s,\,k)}\sum_{k'=1}^{K}\left(U_{i(r,\,k')}V_{i(s,\,k')}W_{i(t,\,k')}\right)\right]}
$$

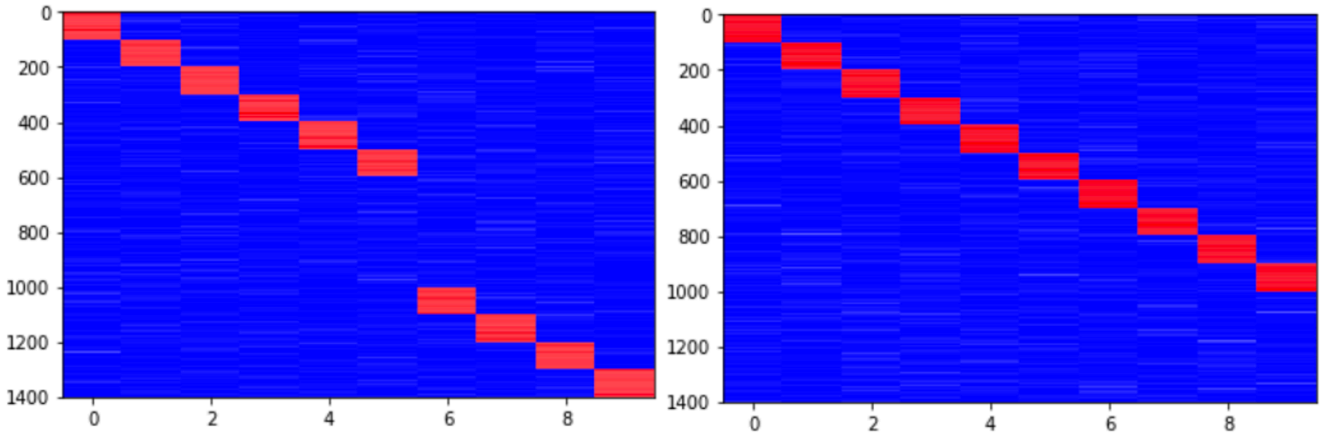
ここで，i は 1 または 2，j は i=1 のとき 2,i=2 のとき 1 である．この式に従い各変数を繰り返し順に更新して最適化を進め，テンソル分解を行うことで目的の行列を得る．

4 実　　験

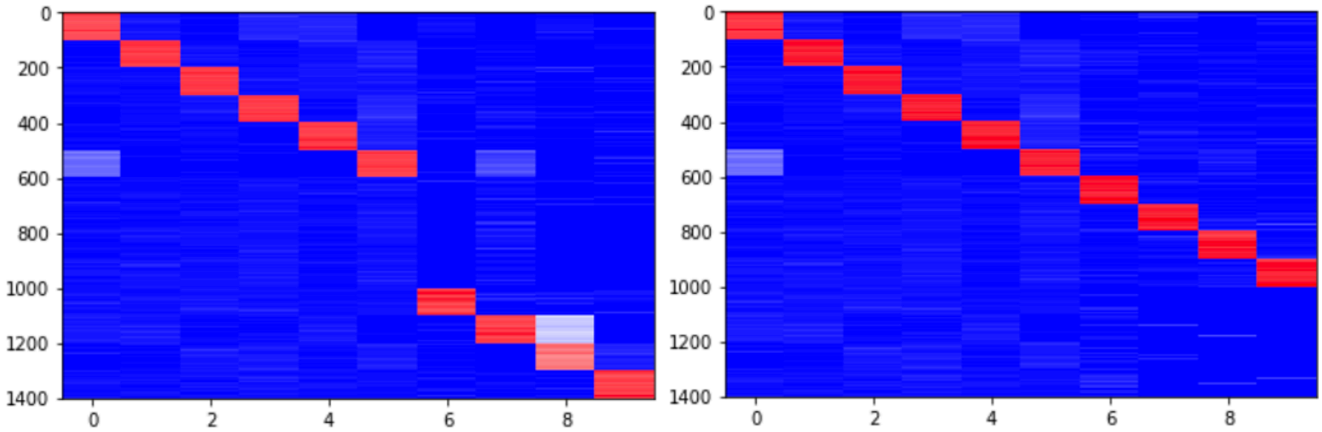
本手法の実験として正解が自明な生成データによる実験と実データである特許データによる実験を行った．

5 生成データによる実験

本節では正解が自明な生成データによる実験について説明する．本実験では，$U_i, V-i, W_i$ のそれぞれの行列を正解データとして生成し，これらの積である X_i を提案手法により行列に分解することで，正解データの行列にどの程度近い結果が得られるかを検証する．トピックのごとの単語分布を示す行列である U_i(単語数 x トピック数) の正解データについて，共通成分と非共通成分が明確に分かれたものにするために，行列 $U_i \in \mathbb{R}_+^{1600\times 300}$ の共通部分と非共通部分の要素を以下のように設定した．



(a) 生成した正解の行列 (左が U_1，右が U_2)



(b) 本手法により再現した行列 (左が U_1，右が U_2)

図 3: 生成データによる実験結果

$$U_{ic(r,\,k)} = \begin{cases} 1 & (100(k-1) < r \leqq 100k) \\ 0 & (otherwise) \end{cases}$$

$$U_{id(r,\,k)} = \begin{cases} 1 & (600 + 400(i\ -\ 1) + 100(r\ -\ 1) < r \\ & \leqq 700 + 400(i\ -\ 1) + 100(k\ -\ 1)) \\ 0 & (otherwise) \end{cases} \quad (5)$$

その後，行列 U_i のすべての要素にガンマ分布に従うランダムな値をノイズとして加えた．文書とトピックの関連度示す行列である V_i と時間とトピックの関連度示す行列である W_i の正解データについては 0〜1 のランダムな値とした．これらの行列の積であるテンソルに提案手法を適応して行列に分解した．ハイパーパラメータである損失関数内の定数 α と β はどちらも 10 に設定して実験を行った．

図 3 が行列 U_i の生成した正解の行列と，本手法の分解により再構築した U_i を可視化したヒートマップである．図 3 に見られるように本手法により再構築した U_i は 10 個のトピックについて，正解データと同じ単語がそれぞれのトピックを表すように抽出されている．一方で，正解データと異なる単語についてもトピックに含まれてしまっている．図 3 の (b) の左図の右から二列目のトピック 8 などは，かなり誤った単語が含まれていることが分かる．提案手法は現段階ではある程度の精度で分解はできるものの，誤まった結果も含まれてしまうといえる．

6 特許データによる実験

6.1 特許データによる実験の結果

本節では建築分野の特許文書データに本手法を適用した実験について述べる．

6.2 特許データによる実験の設定

提案手法を用いて二つの建設会社が出願した特許文書集合を比較する．各企業の特許に共通して出現する技術トピックや各企業の特許に固有の技術トピックを検出し，それらの技術トピックの時間推移によるトレンドの変化を調査する目的があるという想定で実験を行った．日本の特許文書データのうち出願者に大手建設会社である”企業 a”が含まれている特許文書 1650 件と，同じく大手建設会社である”企業 b”が含まれている特許文書 1277 件を分析対象として本手法により比較した．期間は 2001 年 1 月から 2021 年 10 月までに出願された特許文書を用いて，各年ごとの変化を分析する．特許特許文書の，”発明の名称”・”要約”・”詳細な説明”という三つの項目の文に含まれている，名詞が二つ以上連続してできた複合語を対象の用語として tf-idf を算出し三章で説明したデータ構造の 3 階テンソル

企業a	トピック0	トピック1	トピック2	トピック3	トピック4	トピック5	トピック6	トピック7	トピック8	トピック9
上位0	半導体工場	支持基部	天井エリア	熱交換	円環	解泥	増ダイ監視装置	造波板	ベントナイト溶液	防止効果
上位1	固定ナット	レミキサー	臭化ナトリウム	統計データ	水中地盤	入力データ	高圧水供給装置	警報信号	距離データ	枢支軸
上位2	トロール受け	相対回動	圧縮状態	劣化診断	回転ドラム	排砂	スライド運動	底受部移動機構	オーバーフロー	壁体
上位3	安定状態	ウレタン樹脂	土木学会論文集	管理水位	圧力センサー	施工品質	投入工程	磁界発生	タシャフト	帯鉄
上位4	弾性範囲	集水ピット	中高周波数帯域	テールシール	穿孔ロッド	圧接状態	積載荷重	背後地盤	加圧状態	進機自体
上位5	掘削土砂	貫入力	配置状態	硫酸カルシウム	発泡樹脂	シールドジャッキ	三角錐	回転半径寸法	技術的範囲	駆動制御
上位6	風力発電装置	吹付コンクリート	貫通孔	既設護岸	電子メール	液晶パネル	止ねじ	発泡ウレタン	施工装置	供給停止
上位7	梁接合	除塩処理	荷荷重	有機物分解	連結用ボルト	固定セット	変形形態	原水槽	受信装置	所定個所
上位8	所定間隔離間	港湾施設	カッタ駆動モータ	掘削壁面	ビット装着溝	係止凹部	排水経路	浄化処理	主要工程	撤去跡
上位9	土木学会論文集	連結ロープ	略水平	搬送設備	繊維同士	浄化処理	微生物供給	ばね機構	管内圧力	下方斜め

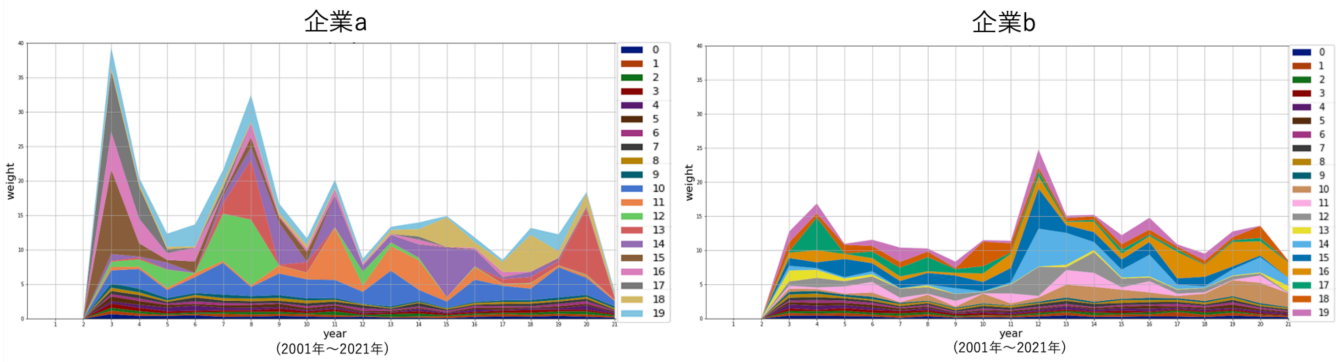
企業a	トピック10	トピック11	トピック12	トピック13	トピック14	トピック15	トピック16	トピック17	トピック18	トピック19
上位0	貫通孔	掘削ビット	電極装置	歩行支援装置	マイクロフォン	クリーンルーム	掘削機本体	破砕装置用電極	支持基板	シールドトンネル
上位1	支持脚	掘削機械	放電破砕	弾性伸縮機構	音圧レベル	プロセスエリア	既設トンネル	破砕装置	掘動	セグメントリング
上位2	地山	油圧ジャッキ	放電破砕装置	コイルばね	音圧波形データ	導電処理	土留壁	外周電極	湾曲凹面	シールドマシン
上位3	基礎梁	掘動基板	電源装置	調整ロープ	画像データ	導電性クロス	エレクタ装置	同軸電極	回転中心	配設
上位4	衝撃音遮断	回転中心	浮遊電極	装置フレーム	映像採取ユニット	ガス発散	シールドジャッキ	中心電極	油圧シリン	地山
上位5	横断面形状	排泥タンク	放電ギャップ	体重負荷	音源推定用画像	シーリング処理	掘削土砂	同軸ケーブル	湾曲凸面	免震
上位6	幅寸法	地山	極端子	左支柱	音採取	天井エリア	掘削壁面	他方ケーブル	円柱状突起	免震装置
上位7	螺着	先頭開口	自由溝	右支柱	周波数帯域	帯電防止効果	既設セグメント	中心導体	発進基地	シールド本体
上位8	平面形状	推進力伝達	圧力伝達媒体	動滑車	データ記憶	石膏ボード	バルクヘッド	放電抵抗	右内面	推進ジャッキ
上位9	継手付き鋼管	集合タンク	放電溝	荷重ばね装置	表示画面	空気洗浄フィルタ	旋回リング	延在	左内面	該立坑

(a) ”企業 a”の特許文書集合から検出された技術トピックの表 (トピック 0〜19)

企業b	トピック0	トピック1	トピック2	トピック3	トピック4	トピック5	トピック6	トピック7	トピック8	トピック9
上位0	半導体工場	支持基部	天井エリア	熱交換	円環	解泥	増ダイ監視装置	造波板	ベントナイト溶液	防止効果
上位1	固定ナット	レミキサー	臭化ナトリウム	統計データ	水中地盤	入力データ	高圧水供給装置	警報信号	距離データ	枢支軸
上位2	トロール受け	相対回動	関連技術	劣化診断	回転ドラム	排砂	スライド運動	底受部移動機構	オーバーフロー	壁体
上位3	安定状態	ウレタン樹脂	土木学会論文集	管理水位	圧力センサー	施工品質	投入工程	磁界発生	タシャフト	帯鉄
上位4	弾性範囲	集水ピット	中高周波数帯域	テールシール	穿孔ロッド	圧接状態	積載荷重	背後地盤	加圧状態	進機自体
上位5	掘削土砂	貫入力	配置状態	硫酸カルシウム	発泡樹脂	シールドジャッキ	三角錐	回転半径寸法	技術的範囲	駆動制御
上位6	風力発電装置	吹付コンクリート	貫通孔	既設護岸	電子メール	液晶パネル	止ねじ	発泡ウレタン	施工装置	供給停止
上位7	梁接合	除塩処理	荷荷重	有機物分解	連結用ボルト	固定セット	変形形態	原水槽	受信装置	所定個所
上位8	所定間隔離間	港湾施設	カッタ駆動モータ	掘削壁面	ビット装着溝	係止凹部	排水経路	浄化処理	主要工程	撤去跡
上位9	分解性材料	連結ロープ	略水平	搬送設備	繊維同士	浄化処理	微生物供給	ばね機構	管内圧力	下方斜め

企業b	トピック10	トピック11	トピック12	トピック13	トピック14	トピック15	トピック16	トピック17	トピック18	トピック19
上位0	製鋼スラグ	鋼管矢板	基礎杭	防水シート	グラブバケット	軟弱地盤	鋼管杭	水シート	鋼矢板	水壁
上位1	ベルトコンベア	圧入装置	捨石マウンド	テールプレート	回動	真空圧密	杭頭部	保護マット	止水ゴム	周期波低減
上位2	潜堤	回転駆動装置	水底地盤	新設防水シート	汚濁防止膜	真空ポンプ	杭支持	土質材料	人工干潟	粘土鉱物
上位3	混合材料	平面形状	透過型海域制御	シールドジャッキ	土砂成分	地盤改良	ブレキャスト	透水係数	潜堤	波工
上位4	混合装置	鋼板セル	止水装置	溶着装置	浚渫土砂	気密性キャップ	コンクリート杭	透水性材料	水底地盤	海成粘土
上位5	一軸圧縮	腹起し	水底基礎	既設防水シート	グラブ浚渫船	排水ホース	装薬孔	重力式岸壁	リート矢板	地下水位
上位6	人工干潟	土留め壁	立設	ガイドレール	浚渫装置	真空圧密工法	水底地盤	水工	水中ポンプ	型鋼矢板
上位7	一軸圧縮試験	石炭灰	杭基礎	接合装置	ラブバケット	レーン工法	斜杭	配設	回動	水護岸
上位8	液性限界	施工場所	水中コンクリート	回動	旋回中心	地盤改良工法	油圧ハンマ	海成粘土	締固め	法線直角
上位9	鉄鋼スラグ	止水性	杭頭	シールド本体	汚濁防止装置	サイフォン機能	ガイドレール	海底地盤	沈埋函	通水口

(b) ”企業 b”の特許文書集合から検出された技術トピックの表 (トピック 0〜19)



(c) 技術トピックの時間推移での増減を示す積み上げ折れ線グラフ

図 4: 特許データによる実験結果
(検出されたトピック 0〜9 は企業間の共通トピック，トピック 10〜19 はそれぞれの企業に固有なトピック)

を作成した．二つ以上連続してできた複合語を対象の用語としたのは，なるべく特許に含まれる技術用語を特徴語とするためである．また対象の用語の選定の際に用語の文書頻度が 5 以上 1000 以下の単語だけを対象にした．ハイパーパラメータである，検出する共通トピックの数 Kc と固有トピックの数 Kd はどちらも 10 として，損失関数内の定数 α と β はどちらも 0.5 に設定した．

本手法を適用した結果，図 4 のような結果が得られた．図 4 の (a) と (b) はそれぞれ ”企業 a ”と ”企業 b ”の特許文書集合から検出された技術トピックを示す表である．表の各列が検出された技術トピックを表しており，これはトピックを表す単語分布から重みが上位 10 番以内の用語を上から並べたものである．(a) と (b) の表の左半分に含まれるトピック 0〜9 は二つの企業の特許文書集合に共通して含まれているとして検出された技術トピックであり，右半分に含まれるトピック 0〜10 はそれぞれの企業に固有の成分として検出されたトピックであ

る．図 4 の (c) は (a) と (b) に示された各トピックが時間推移でどのように増減しているかを示す積み上げ折れ線グラフである．左のグラフが”企業 b”について，右のグラフが”企業 a”についてのものである．横軸が時間であり各メモリが 2001 年から 2021 年の各年を表す．縦軸はトピック 0〜19 がそれぞれの時期にどれくらい文書に含まれていたを表す重みである．下から順にトピック 0〜19 の値を積み上げて年ごとのトピックの重みを可視化しており，グラフの下から 10 トピック (0〜9) は両グラフの共通トピックである．

6.3 特許データによる実験の結果の考察

実験結果からそれぞれの企業が強みとしている技術分野とそのトレンドを知ることができる．例えば”企業 a”の固有技術トピックであるトピック 13 は”歩行支援装置”・”体重負荷”などの用語からなる技術トピックで，ここから”企業 a”が歩行支援器具の技術を持っていることが分かる．”企業 a”のグラフを見るとこのトピック 30 は 2007 年から 2008 年と 2019 年や 2020 年に出願された特許に多く含まれていたことも知ることができる．

一方で，企業間に共通の技術トピックであるトピック 0〜9 は上位の用語を見てもあまりまとまったトピックとして検出されていない．”企業 a”と ”企業 b ”の間で共通している技術がなかったために，この結果になっている可能性もあるが，提案手法の問題や提案手法とデータの相性なども考えられる．たとえば，企業ごとに用語の使い方に癖がある場合，同じ技術にも関わらず別の技術トピックとして分類されている可能性なども考えられる．問題の所在の特定と改善のためにさらなる実験・考察が必要であるといえる．

7 結　　論

本論文では Non-negative Tensor Factorization(NTF) [10] と Joint Non-negative Matrix Factorization (Joint-NMF) [5] の二つの手法を組み合わせて発展させた，二つの時系列文書集合間で共通するトピックと各文書集合に固有のトピックとそれらの時間推移を同時に発見する新しい手法を提案した．生成データによる実験では提案手法により共通トピックと非共通トピックをある程度の精度で検出できていることを確認した．実データである特許文書による実験では，本手法を用いて得られた共通・非共通トピックと時系列グラフを見ることで企業が取り組んでいる技術や取り組んでいた時期などについて文書を読まずとも把握できることを示した．

一方で，結果に誤差が含まれることや実データ (特許データ) において共通トピックが検出されづらい可能性があるという問題も明らかになった．今後の課題として，第一に生成データによる実験結果に見られたような誤差を減らす方法の検討がある．Joint Non-negative Matrix Factorization(Joint-NMF) [5] の論文で提案されている pseudo-deflation method などといった手法を提案手法に応用することで精度が向上する可能性がある．第二に特許データの実験結果のように実データにおいて共通成分があまり抽出されないことについて検討する必要がある．本当に問題なのか，問題であるなら手法に原因があるのか，実データとの相性にあるのか，など，問題を特定するために実験を行いその上で改善を進める必要がある．

謝　　辞

本研究の一部は JST CREST (JPMJCR16E3)，JSPS 科研費（JP21H03552，JP20K23337） および株式会社熊谷組の支援による．

文　　献

[1] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120, 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] Sunil Kumar Gupta, Dinh Phung, Brett Adams, and Svetha Venkatesh. Regularized nonnegative shared subspace learning. *Data mining and knowledge discovery*, Vol. 26, No. 1, pp. 57–97, 2013.

[4] Tamir Hazan, Simon Polak, and Amnon Shashua. Sparse image coding using a 3d non-negative tensor factorization. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, Vol. 1, pp. 50–57. IEEE, 2005.

[5] Hannah Kim, Jaegul Choo, Jingu Kim, Chandan K. Reddy, and Haesun Park. Simultaneous discovery of common and discriminative topics via joint nonnegative matrix factorization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, p. 567–576, New York, NY, USA, 2015. Association for Computing Machinery.

[6] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, Vol. 401, No. 6755, pp. 788–791, 1999.

[7] Chunshan Li, William K Cheung, Yunming Ye, Xiaofeng Zhang, Dianhui Chu, and Xin Li. The author-topic-community model for author interest profiling and community discovery. *Knowledge and Information Systems*, Vol. 44, No. 2, pp. 359–383, 2015.

[8] Yasuko Matsubara, Yasushi Sakurai, and Christos Faloutsos. Non-linear mining of competing local activities. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 737–747, 2016.

[9] Bree Nordenson. Overload. *Columbia journalism review*, Vol. 47, No. 4, pp. 30–42, 2008.

[10] Amnon Shashua and Tamir Hazan. Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd international conference on Machine learning*, pp. 792–799, 2005.

[11] Ajit P Singh and Geoffrey J Gordon. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 650–658, 2008.

[12] Carmen K. Vaca, Amin Mantrach, Alejandro Jaimes, and Marco Saerens. A time-based collective factorization for topic discovery and monitoring in news. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, p. 527–538, New York, NY, USA, 2014. Association for Computing Machinery.