Group Sparse NMF に基づくグラフ構造推定法の提案

情報数理応用研究

5217C009-9　河部瞭太
指導教員　後藤正幸

A Proposal of Graph Structural Estimation Model Based on Group Sparse NMF

KAWABE Ryota

1　研究背景・目的

近年，情報技術の発展および SNS（Social Networking Service）等の広まりに伴い，様々な主体のつながりを表現するデータが膨大に蓄積されるようになっている．これらのデータは全体としてグラフで表現できることから，様々な分野において大規模なグラフ構造データからの知識発見を目的としたグラフマイニング手法の研究が進められている．例えば，SNS におけるユーザ同士の友人関係はノードおよびそれらの間のリンクを用いた大規模かつスパースなネットワーク（グラフ）として表現される．一般に SNS は大量のユーザが利用しているが，個々のユーザは数十から数百程度の友人関係しか持っていない．したがってネットワークを構成するノード数に対しリンク数は非常に少なくなり，ネットワークはスパースであることが多い．そのようなグラフ構造はソーシャルグラフと呼ばれ，その活用のために多くの研究がなされている．

一方，インターネット上の多数の Web サイト間の関係もグラフ構造で捉えることができ，有用な知見を抽出できる対象データの一つである．従来，これらの Web サイト間の構造分析は「互いのリンクのあり／なし」などの物理関係によって作られたネットワークの構造に対してグラフマイニングの手法が適用されていた．しかし，現在では多数のユーザによる大規模な Web 閲覧履歴データが蓄積できるようになっており，こういったデータは Web サイトの関係分析に大変有用であると考えられる．一般に，Web サイトの閲覧履歴データ（アクセスログ）は，ユーザがどのサイトを閲覧したかを時系列順に並べたものとなっており，ユーザ属性情報や閲覧行動の分析を行うことによってユーザの嗜好の把握やマーケティング施策の最適化に結びつけることができる．このユーザ閲覧行動という観点から Web サイト間の関係性や構造を分析することで，潜在的な特徴を明らかにすることが期待されるようになってきている．アクセスログは他にも，サイトの特性や閲覧行動の間の関係性を分析を通じて Web サイトの高品質化や広告最適化に活用可能である．そこで，Web サイト間の関係を，ユーザの閲覧行動によって定義されるグラフ構造として捉え，アクセスログに対してグラフマイニング手法を適用することを考える．

グラフマイニング手法を適用するためには，ユーザの閲覧履歴データを Web サイト間のグラフ構造データに変換する必要がある．ユーザ数や閲覧回数が膨大なため，関係の有無を簡単に 2 値で定義するような，グラフ構造へ変換するための確立された方法は存在しない．最も簡単な方法として，共通に閲覧しているユーザ数の大小で関係性の有無を定義することが考えられるが，分析者による恣意的な閾値の設定が必要になってしまう．異なる方法として，2 つの Web サイトを両方閲覧したユーザが一人でもいる場合に，これらの Web サイトは関係があるとしてリンク有りのネットワークを生成する方法も考えられる．この場合ほとんどの Web サイト間でリンクが繋がった密なネットワークになってしまい，分析の価値がなくなってしまう．さらに，これらの方法では，2 つのサイト間の共起回数という局所的な情報だけでリンクの有無を決めてしまい，他サイトも含めた Web 全体に対するユーザの閲覧傾向を考慮することができていない．このように統計的な構造と Web サイト全体の関係性を考慮しつつ，ユーザによる大規模閲覧履歴を集約したネットワークを生成するためには，新たなモデルの構築が望まれる．そこで，各ユーザのネットワーク上の対象 Web サイト全体の閲覧行動データを考慮しながら，ネットワーク構造を生成する手法の提案を本研究の目的とする．

そのための具体的な方法として，本研究では，閲覧データに対してその統計的構造を保持しつつ，潜在的な特徴を考慮しながら 2 つの行列へ分解する手法である，NMF（Non-negative Matrix Factorization）[1] の導入を考える．しかしながら，NMF で潜在特徴について分解した内部表現を Web サイト間の類似度計算にそのまま用いた場合，ほとんどのノード間にリンクが引かれることになりグラフマイニング手法による分析の価値がなくなってしまう．したがってネットワーク内のリンクが過多にならないようにサイトの特徴ベクトルを直交させ，スパースなネットワークを獲得する必要がある．スパースな潜在特徴ベクトルが得られるような分析方法としては，NMF にスパース性を取り入れた L_1NMF[2] が存在する．しかしこの方法では，観測行列を分解して得られたサイト－潜在特徴行列中のいくつかの要素が 0 になってもベクトル同士の類似度が 0 になることを保証していない．類似度そのものを正則化することは困難であるため，何らかの方法で不必要な類似度が 0 になるような方法を考える必要がある．そこで，本論文ではグループごとに 0 か非 0 を選択する Group Lasso[3],[4] を NMF に適用することにより，表現能力を落とさずリンクの数を減らす手法を提案する．

さらに本研究では，インターネット調査会社，株式会社ヴァリューズより提供された Web アクセスログデータに提案手法を適用する．この実データ分析を通して提案手法で作成した類似度行列の有効性を実証し，得られた結果を用いた Web サイト分析を行う．

2　準備

2.1　NMF

NMF（Non-negative Matrix Factorization）は非負行列 $\boldsymbol{A}$ を潜在特徴を考慮しながら 2 つの非負行列 $\boldsymbol{X}$ と $\boldsymbol{W}$ に分解する次元圧縮手法の 1 つである．

$$\boldsymbol{A} \approx \boldsymbol{X}\boldsymbol{W} \tag{1}$$

ここで Web サイトの集合を $\mathcal{S}=\{S_i : 1 \leq i \leq I\}$，ユーザの集合を $\mathcal{U}=\left\{U_j : 1 \leq j \leq J\right\}$ と定義する．ユーザ U_j が Web サイト S_i を見たかどうかを $a_{ij} \in \{0,1\}$ で表すと，ユーザと Web サイトの閲覧行列は $A=\left[a_{ij}\right] \in \mathbb{R}^{I \times J}$ と定義できる．潜在特徴空間を K 次元とするとユーザと潜在特徴に関する行列は $\boldsymbol{X}=[x_{ik}] \in \mathbb{R}^{I \times K}$，Web サイトと潜在特徴に関する行列は $\boldsymbol{W}=\left[w_{kj}\right] \in \mathbb{R}^{K \times J}$ のように表すことができる．

$$\begin{aligned} &\min_{\boldsymbol{X},\boldsymbol{W}} ||\boldsymbol{A}-\boldsymbol{X}\boldsymbol{W}||_F^2 \\ &s.t. \quad \forall x_{ik} \geq 0, \forall w_{kj} \geq 0 \end{aligned} \tag{2}$$

ここで $||\cdot||_F$ はフロベニウスノルムを意味する．

この最適解は，式 (3)-(4) を繰り返し用いて更新し，収

束させることで求めることができる．t 回目の更新で得られた $\boldsymbol{X}^{(t)}$ と $\boldsymbol{W}^{(t)}$ の各成分を $x_{ik}^{(t)}$，$w_{kj}^{(t)}$ と定義すると，これらの更新は次式で与えられる．

$$x_{ik}^{(t)} = x_{ik}^{(t-1)} \frac{\left(\boldsymbol{A}^{(\mathbf{t-1})}\boldsymbol{W}^{(\mathbf{t-1})\boldsymbol{T}}\right)_{ik}}{\left(\boldsymbol{X}^{(\mathbf{t-1})}\boldsymbol{W}^{(\mathbf{t-1})}\boldsymbol{W}^{(\mathbf{t-1})\boldsymbol{T}}\right)_{ik}} \tag{3}$$

$$w_{kj}^{(t)} = w_{kj}^{(t-1)} \frac{\left(\boldsymbol{X}^{(\mathbf{t-1})\boldsymbol{T}}\boldsymbol{A}^{(\mathbf{t-1})}\right)_{kj}}{\left(\boldsymbol{X}^{(\mathbf{t-1})\boldsymbol{T}}\boldsymbol{X}^{(\mathbf{t-1})}\boldsymbol{W}^{(\mathbf{t-1})}\right)_{kj}} \tag{4}$$

2.2 Lasso

Lasso（Least absolute shrinkage and selection operator）はデータ分析において線形モデルを推定する際の回帰係数の絶対値の和 (L_1 ノルム) を正則化項としてモデルを推定する手法である．回帰係数の一部を完全に 0 と推定することから，パラメータの推定と変数選択を同時に実行できる手法として注目を集めている．

係数ベクトル $\boldsymbol{w}$ の L_1 ノルムは各要素の絶対値の線形和として以下のように定義される．

$$||\boldsymbol{w}||_1 = \sum_{j=1}^{d} |w_j| \tag{5}$$

ただし，$||\cdot||_1$ は L_1 ノルムを表し，非ゼロの要素が増えるほど L_1 ノルムは大きくなる．$\hat{L}(\mathbf{w})$ を最適化する関数とした時以下の問題を解くことになる．ここで $\lambda\,(>0)$ は正則化パラメータである．

$$\min_{\boldsymbol{w}} \hat{L}(\boldsymbol{w}) + \lambda||\boldsymbol{w}||_1 \tag{6}$$

2.3 Group Lasso

あらかじめ定義したグループ単位で変数をゼロもしくは非ゼロに選択する回帰手法として Group Lasso が提案されている [3]．d 個の変数の適当な分割を $\mathfrak{G}$ と定義する．例えば d を偶数として，$\mathfrak{G} = \{\{w_1, w_2\}, \{w_3, w_4\}, ..., \{w_{d-1}, w_d\}\}$ は 2 つずつの変数の組への分割を表す．ここで，L_p ノルムに基づく係数行列 $\boldsymbol{w}$ のグループ L_1 ノルム正則化項は以下のように定義される．

$$||\boldsymbol{w}||_{\mathfrak{G}} = \sum_{\mathfrak{g}\in\mathfrak{G}} ||\boldsymbol{w}_{\mathfrak{g}}||_p \tag{7}$$

ここで $\boldsymbol{w}_{\mathfrak{g}}$ は分割された一つの列 $\mathfrak{g}$ に対応する $|\mathfrak{g}|$ 次元ベクトルを表す．グループの大きさがすべて同じ場合 $d' = \frac{d}{|\mathfrak{G}|}$ として行列 $\boldsymbol{W} = [\boldsymbol{w}_1, ..., \boldsymbol{w}_{|\mathfrak{G}|}]^T \in \mathbb{R}^{|\mathfrak{G}|\,\times\,d'}$ を定義するとブロック p,q ノルムは以下のように表せる．

$$||\boldsymbol{W}||_{p,q} = \left(\sum_{j=1}^{|\mathfrak{G}|} ||\boldsymbol{W}_{j,:}||_p^q\right)^{\frac{1}{q}} \tag{8}$$

ここで $\boldsymbol{W}_{j,:}$ は行列 $\boldsymbol{W}$ の第 j 行ベクトルを表す．一般的に $p=2$ としたブロック 2,1 ノルムがよく用いられるため，本論文ではこれ以降ブロック 2,1 ノルムを用いる．

Group Lasso ではグループの構成は事前に定めておく必要があるが，グループ単位でゼロか非ゼロとなるようにパラメータ推定したい場合に有効な手法である．

3 提案手法

3.1 提案の着想

グラフマイニング手法は主にソーシャルグラフと呼ばれるノードとリンクから構成されるスパースな大規模ネットワークに適用されることが多い．これに対し，本論文では Web サイトの閲覧行動や商品購買行動のようなログデータからスパースなグラフ構造データへと変換し，グラフマイニング手法を用いた分析を適用可能とするための方法を提案する．

そのためにまず，Web サイトにおけるユーザの閲覧行動や購買行動から，Web サイトやアイテム間のグラフ構造を構成する必要がある．しかし，ユーザの行動履歴を適切に反映したグラフ構造データの作成については，現在のところ確立した方法は存在しない．例えばコサイン類似度を用いて Web サイト間のリンクの重みを計算した場合，閲覧行動がスパースであるためほとんどの Web サイト間の類似度が 0 になってしまい，他のサイトとリンクが引かれない孤立ノードが発生してしまう．また，単にそれら 2 つのサイトを共に見ているユーザがいるか否かしか考慮できず閲覧全体の傾向を反映できない．そこでユーザごとのサイト全体の閲覧履歴を潜在特徴を使った手法で圧縮する方法が望ましいと考えられる．

ここで潜在特徴を考慮しながら次元圧縮を行う手法として，NMF が知られている．NMF を適用してグラフ構造を作成する場合，次元圧縮した特徴空間で Web サイトをベクトル表現し，これらの類似度を測り，その類似度をリンクの重みとすることが考えられる．しかし，多くのサイト間の類似度が非ゼロになってしまい，グラフとして表した際リンクが多すぎて把握が困難になることや，サイト間の関係の分析が困難となったり，クラスタリングがうまくできなくなったりすることが考えられる．ここで，正則化項を導入した NMF である L_1NMF が提案されているが，もともとグラフ生成を意図した手法ではないため，潜在特徴ベクトルの要素を 0 にすることはできてもそれらの間の類似度のほとんどは 0 にならない．一方で，特徴ベクトルが直交し類似度が 0 になるまでスパース性を高くした場合，分解後の行列のほとんどの要素が 0 になってしまいモデルとしての表現能力が落ちてしまう．そのため，類似度を適切に 0 にすることを目的とした手法が必要となるが，類似度そのものを正則化することは困難である．

そこで，本研究では，近年機械学習の分野で注目されている Group Lasso を NMF に適用し，グループごとにまとめて潜在特徴を 0 にする方法を提案する．これにより，各 Web サイトの特徴ベクトル同士が直交しやすくなることで，微小な類似度を適切に 0 にすることが可能になる．そして，求められたスパースな特徴行列をもとに適切なグラフを描写すると同時に，従来の研究で仮定されていたようなグラフ構造データへの変換をすることで，これまで数多く提案されてきたグラフマイニング技術の適用を可能とする．

3.2 定式化

本研究では，NMF に Group Lasso のアプローチを適用した方法を提案する．これは，観測行列 $\boldsymbol{A}$ を分解する際に，サイト特徴行列 $\boldsymbol{W}$ の推定に対して Group Lasso のアプローチを適用するものである．ここで，$\boldsymbol{X}$ と $\boldsymbol{W}$ を同時に推定することはできないため，一方を更新する際は他方を固定の説明変数として扱い交互に更新を行う．

ここで，Group Lasso は，正則化項が微分不可能であるため，直接解を求めることができない．そのため，探索的に推定する手法として微分可能な誤差関数に対する最も一般的な最適化法である近接勾配法を適用する．この手法では微分可能な損失項と微分不可能な正則化項を区別して扱う．最適化関数は以下のようになる．

$$f(\boldsymbol{w}) = \hat{L}(\boldsymbol{w}) + \lambda||\boldsymbol{w}||_{\mathfrak{G}} \tag{9}$$

ここで L_1 正則化において最適化問題を最小化するための prox 作用素が一般的に用いられている．NMF の更新において Web サイト－潜在特徴に対する行列 $\boldsymbol{W}$ の更新を Group Lasso に対応した prox 作用素を用いた加速付き勾配法 [5] に基づき行う．以下に Group Lasso に用いる prox 作用素を表す．

$$\mathrm{prox}_{\lambda}^{\mathfrak{G}}(\boldsymbol{y}) = \underset{\boldsymbol{w}\in\mathbb{R}^d}{\arg\min}\left(\frac{1}{2}||\boldsymbol{y}-\boldsymbol{w}||_2^2 + \lambda\sum_{\mathfrak{g}\in\mathfrak{G}}||\boldsymbol{w}_{\mathfrak{g}}||_2\right) \tag{10}$$

これは解析的に以下のように書き表される．

$$\begin{aligned}[\mathrm{prox}_{\lambda}^{\mathfrak{G}}(\boldsymbol{y})]_{\mathfrak{g}} &= \mathrm{prox}_{\lambda||\cdot||_2}(\boldsymbol{y}_{\mathfrak{g}}) \\ &= \begin{cases}(||\boldsymbol{y}||_2-\lambda)\frac{\boldsymbol{y}}{||\boldsymbol{y}||_2} & if\ ||\boldsymbol{y}||_2>\lambda, \\ \boldsymbol{0} & otherwise.\end{cases}\end{aligned} \tag{11}$$

ここで $[\cdot]_{\mathfrak{g}}$ は行列内のグループ $\mathfrak{g}$ を表す．

しかし，そのまま加速付き勾配法の計算を行うと更新した際に負の値が生じてしまい NMF の非負の前提に反してしまうため，更新ごとに負値を 0 に置換する．また，初期値として L_1NMF の結果を用いる．$\boldsymbol{W}$ の更新のたび，通常の NMF と同様の更新式を用いて $\boldsymbol{X}$ を更新する．

3.3 提案手法のアルゴリズム

潜在特徴変数に対して適切な分割が与えられていることを前提とし，提案手法のアルゴリズムを以下に示す．

提案手法のアルゴリズム

1. 適切に $\boldsymbol{W}^0$ と学習率 η_1 を初期化し，$\boldsymbol{Z^1}=\boldsymbol{W^0}$, 加速法における加速のパラメータ $s_0=1$，$t=1$ とする. 行列の初期値として L_1NMF の結果を用いる．
2. 収束するまで以下を繰り返す (t 回目の更新).

a)$\boldsymbol{W^{t+1}}$ の更新:

$$\boldsymbol{W^{t+1}} = \mathrm{prox}_{\lambda\eta_t}^{\mathfrak{G}}\left(\boldsymbol{Z^t} - \eta_t\Delta\hat{L}\left(\boldsymbol{Z^t}\right)\right) \tag{12}$$

ここで負の要素が発生した時，その要素ごとに 0 に置換する．最適な η_t は数式的に求めることができないため，ここでバックトラッキング法を用い適切な値を探索する．

b)$\boldsymbol{Z^{t+1}}$ の更新:

$$\boldsymbol{Z^{t+1}} = \boldsymbol{W^{t+1}} + \left(\frac{s_{t-1}}{s_t-1}\right)(\boldsymbol{W^{t+1}}-\boldsymbol{W^t}) \tag{13}$$

ここで $s_{t+1}=(1+\sqrt{1+4s_t^2})/2$.

c)$\boldsymbol{X^{t+1}}$ の更新:

$$\boldsymbol{X^{(t+1)}} = \boldsymbol{X^{(t)}}\frac{\left(\boldsymbol{A}\boldsymbol{W^{(t+1)T}}\right)}{\left(\boldsymbol{X^{(t)}}\boldsymbol{W^{(t+1)}}\boldsymbol{W^{(t+1)T}}\right)} \tag{14}$$

ただし，η_t は H-平滑の時 $\frac{1}{H}$ とするのが最も良いが，数式的に求めることができないため，バックトラッキング法を用いて収縮させながら探索する．バックトラッキング法のアルゴリズムを以下に示す．

バックトラッキング法のアルゴリズム

1. 適切に η_t を初期化する．
2. 収束するまで繰り返す:
   a) 停止条件 (15) を満たせば終了:

$$\begin{aligned}\boldsymbol{W^{t+1}} \le \boldsymbol{f}\left(\boldsymbol{Z^t}\right) + \langle\boldsymbol{\Delta}\hat{L}\left(\boldsymbol{Z^t}\right), \boldsymbol{W^{t+1}}-\boldsymbol{Z^t}\rangle \\ + \frac{1}{2\eta_t}||\boldsymbol{W^{t+1}}-\boldsymbol{Z^t}||_2^2\end{aligned} \tag{15}$$

ただし $\langle\cdot\rangle$ は内積を表す．

b) 停止条件を満たしていなければ次式で更新する:

$$\eta_{t+1} = \beta\eta_t\ (0<\beta<1) \tag{16}$$

4 提案手法を用いた実データ分析

実験データとして株式会社ヴァリューズより提供された Web サイト閲覧ログデータを用いる．観測データは 2017/08/01 から 2017/10/31 の期間に収集されたもので，ユーザ数は 1,000 人，対象 Web サイトは閲覧数が 1,000 以下のものから降順に 200 件を選択した．サイト－ユーザ行列を入力データとし，観測行列中の各要素は，あるユーザがそのサイトにアクセスしていれば 1，していなければ 0 とする．

本実験では，提案の妥当性を示すため，与えられた閲覧データを適切にスパースなグラフへと変換することができていることを確認した上で，それが従来から用いられているグラフマイニング技術に適用可能であることを示す．実験手順を以下に示す．

本提案における実験手順

1. サイト－ユーザの閲覧行列に提案手法を適用し，サイト－潜在特徴行列を得る．
2. 得られた行列を用いてすべての Web サイト間のコサイン類似度を計算する．
3. 類似度行列を用いてグラフクラスタリングを行う．

ここではクラスタリング手法としてモデュラリティが大きくなるように自動的にクラスタ数を決定する louvain アルゴリズム [6] を用いる．モデュラリティ値は 0 から 1 の値を取り，1 に近ければ近いほどよい分割であることを示す指標である．サイト a とサイト b の類似度を s_{ab}，クラスタリングにより分割されたコミュニティの集合を $\mathcal{C}=\{u,v,\ldots\}$，グラフ内の重みの総和を m とすると，モデュラリティ値 Q は以下の式で表される．

$$Q = \sum_{u,v\in\mathfrak{C}}\left(\frac{\sum_{a\prime b\prime\in u}s_{a\prime b\prime}}{2m} - \left(\frac{\sum_{a\prime\in u}\sum_{b\prime\in v}s_{a\prime b\prime}}{2m}\right)^2\right) \tag{17}$$

この分析では提案手法の潜在特徴は 15 とし，3 つずつの特徴が 5 グループ存在すると仮定する．$\lambda=9$ とし，NMF および L_1NMF では潜在特徴を $k=12, 13, 14, 15$ として実験を行った．

まず表 1 では提案手法の目的が達成できているか示すため L_1NMF と提案手法により得られた類似度行列において要素が 0 である要素の割合を示す．また，表 2 に NMF，L_1NMF，提案手法のそれぞれで得られた類似度行列に対してクラスタリングを行った際のモデュラリティ値とクラスタ数の比較を示す．

表 1: 提案手法と L_1NMF の 0 の割合の比較

	L_1NMF				提案
潜在特徴の数	15	14	13	12	15
0 の要素の割合（対角除く）	16.5%	15.9%	15.8%	14.1%	39.3%

提案手法より得られた類似度行列は L_1NMF に比べ 0 要素の割合が大幅に増加しており，目的としていたリンク数の減少が可能となったことがわかる．またその行列を用いてクラスタリングを行った際にモデュラリティ値が最も良くなった．これらより提案手法の有用性を示すことができた．

表 2: 提案手法と比較手法のモデュラリティ値と 0 率

	NMF				L_1NMF				提案
潜在特徴の数	15	14	13	12	15	14	13	12	15
クラスタ数	4	3	4	3	3	3	3	3	4
モデュラリティ値 Q	0.298	0.317	0.321	0.298	0.329	0.321	0.320	0.330	0.346

表 3: クラスタごとのユーザ属性の比率

クラスタ	性別		年齢					
	男性	女性	30 代	40 代	50 代	60 代	70 代	80 以上
1	75.2%	24.8%	6.7%	15.5%	39.1%	27.0%	9.3%	2.3%
2	70.3%	29.7%	1.2%	10.2%	33.3%	**34.9%**	**17.7%**	2.7%
3	72.7%	27.3%	**9.4%**	**17.7%**	43.1%	22.1%	6.0%	1.5%
4	**59.7%**	**40.3%**	3.3%	13.4%	41.0%	28.9%	10.7%	2.6%

次に，提案手法で作成した類似度行列を用いてクラスタリングを行い，得られたクラスタについてサイトやユーザの属性の点から傾向を分析し，提案手法の有効性を示す．ここで，データではサイトのカテゴリ（グーグルは“検索エンジン”である等）がいくつかのサイトについて定義されている．サイトカテゴリは大カテゴリとして 25 種，小カテゴリとして 109 種存在するが，頻繁に登場するカテゴリは一部に偏っている．また，ユーザの属性情報については年齢，性別，所在，未婚既婚や子供の有無などが含まれている．

得られたクラスタは 4 つであり，それぞれのクラスタに対するサイトの所属数の割合はクラスタ 1 は 32.7%，クラスタ 2 は 31.1%，クラスタ 3 は 22.2%，クラスタ 4 は 13.9%であった．サイトの所属カテゴリの内訳を見てみると，クラスタ 1 では“製品・サービス”のサイトが多く，クラスタ 2 では“クーポン・ポイント”サイトが多く，クラスタ 3 は“ニュース”サイトが多く，クラスタ 4 は多種のカテゴリのサイトが混在しているということがわかった．これによりクラスタごとにサイトの傾向が分かれることが確かめられた．

一方，表 3 にそれぞれのクラスタを閲覧したユーザ属性のうち，特徴的な結果が得られた性別と年齢について内訳を示した．内訳を求める際，まずそれぞれのサイトごとに比率を求め，クラスタごとにそれらの平均をとって全体の比率を求めた．表 3 より，クラスタ 4 では女性の比率が高く，クラスタ 2 では年齢層が高く，クラスタ 3 では比較的年齢層が低いといったことがわかる．実際，“クーポン・ポイント”サイトは高齢女性に多く利用されていることからも結果の妥当性を示すことができる．

5 考察

4 章で示した結果から提案手法によって変換されたスパースなグラフ構造データは従来手法に比べてモデュラリティ値が高く，グラフクラスタリングにおいて有用であることがわかった．これらの結果から対象セグメントやサイト間の類似度を考慮した広告を打つなどのマーケティングを行うことができる．

また，類似度行列における 0 要素の割合はパラメータ λ に依存する．グラフを視覚的に把握しやすくするためには λ を大きくする必要があり，目的に応じて調整が必要である．

また，本実験では行列計算に与える L_1NMF の結果を求めるための初期値としてランダムな値を与えて実験を行った．収束する際の損失関数の値は同程度であったが，そのたびに求めた行列おける非ゼロのグループの位置は変わっていた．したがって目的に応じた適切な初期値の設定については検討の余地がある．

6 まとめと今後の課題

本研究ではグループ単位のスパース性を導入した NMF による類似度行列を用いたグラフクラスタリングの手法を提案した．提案手法によってノード間のリンクがスパースになるような類似度行列を作成するための，グループ単位で 0 か非 0 になる潜在特徴の表現が可能になった．提案手法によって作成された類似度行列を用いることによって，これまで適用することが難しかったグラフマイニング手法の活用が可能になった．

また，提案の妥当性を示すために実データを提案モデルに基づきグラフ構造へと変換し，グラフクラスタリングを行った．その結果，提案手法を用いて作成された類似度行列は比較手法に比べ良い性能を示した．また，実際にクラスタリングされた結果に基づき，クラスタごとに所属する Web サイトや閲覧しているユーザの傾向に違いがあることを確認し，提案手法の有効性を示すことができた．

今後の課題としては，グループの分け方についての検討が挙げられる．本研究では順に d 個ずつのまとまりでグループを構成しているが，これが最適であるという保証はないため，さらに良い分割方法について研究の余地がある．また，パラメータ λ はいくつかのパターンを試して最適となったものを用いたが，要求されるエッジ数に応じた自動的な定め方を考案する必要がある．本研究では提案手法で作成したグラフ構造をグラフクラスタリングに適用したが，他のグラフマイニング手法にも適用し有効性を示すことが望まれる．

参考文献

[1] D.D. Lee and H.S. Seung, “Algorithms for nonnegative matrix factorization,” *in Advances in Neural Information Processing Systems 13*, pp. 556−562, 2000.

[2] Hoyer, P.O, “Non-negative Matrix Factorization with Sparseness Constraints,” *Journal of Machine Learning Research*, Vol. 5, pp. 1457−1469, 2004.

[3] M. Yuan, Y. Lin, “Model selection and estimation in regression with grouped variables,” *J. Roy. Stat. Soc. B*, Vol. 68, No. 1, pp.49-67, 2006.

[4] 富岡亮太,「スパース性に基づく機械学習,」 講談社, 2015.

[5] P.L. Combettes, V.R. Wajs, “Signal recovery by proximal forward-backward splitting,” *Multiscale Model. Simul.*, Vol. 4, No. 4, pp.1168-1200, 2005.

[6] Blondel V D, Guillaume J-L, Lambiotte R and Lefebvre E, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *J. Stat. Mech.* , P10008, 2008.