ユーザの行動履歴データを活用したネットワーク分析

後藤　正幸

本稿では，顧客の購買履歴から商品間の関係性をネットワーク分析したり，ユーザの閲覧履歴データから Web サイト間の関係性をネットワーク分析するような課題を対象とし，“単純にノード間の関係性の有無が定義できない問題”に対して，グラフマイニング手法を適用する方法について紹介する．そのためにまず，一般的な問題の記述を行うとともに，k-NN グラフと ε-NN グラフというグラフデータ化の手法とそのための類似度計算の手法を解説する．さらに，ユーザの閲覧履歴データに基づいて Web サイト間のネットワーク分析を行った事例を紹介し，NMF を用いて次元圧縮をした後にグラフマイニング手法を適用することで，きれいなコミュニティが抽出できるようになるケースが存在することを示す．

キーワード：ネットワーク分析，k-NN ネットワーク，閲覧履歴，グラフマイニング，クラスタリング

1. はじめに

近年，Social Networking Service (SNS) などのソーシャルメディアの普及に伴い，データサイエンスが対象とするデータ構造は，従来の表構造のものから，人・もの・場所などの多様な対象同士のつながりを表現するグラフ構造へと広がっている [1]．通常，グラフ構造やネットワーク構造を有したデータとは，ノードを表す対象同士の物理的に明確な関係性を指していることが多い．グラフ構造を有したデータの事例としては，Web サイト間のハイパーリンク構造や SNS サイトにおける友達関係やフォロー関係の構造，学術論文間の引用関係の構造などがよく知られているが，これらは関係の有無が明確に定義されやすい対象に対して，その有無によってノード間のリンクが引かれるグラフデータの事例である．たとえば，著名人を紹介した Wikipedia のページ間のハイパーリンク構造は，リンクの有無によってノード（Wikipedia のページ）間の関係性の有無が明確に定義できる．このリンクの有無によって，著名人間の関係性をグラフ化することで，コミュニティを発見するためのクラスタ分析やページランク分析などのグラフマイニング手法を適用することができ，仲間のグループを見つけたり，影響力のある人を見つけることが可能となる [1, 2]．

一方，マーケティング分野などで頻出する購買関係や閲覧関係なども“さまざまなつながりを表すデータ”とみなすことができ，グラフ構造データとして分析することができる．たとえば，「ある顧客がある商品を買った／買わない」，「あるユーザがある映画を見た／見ない」といった関係の有無でグラフ構造を生成し，グラフマイニング手法を適用することは，しばしば有用な知見を導いてくれる．実世界に存在するネットワークでは，多くの事例において，密に繋がったノード群からなるコミュニティが随所に存在することが知られており [3]，このような構造を利用したネットワーク分析手法の適用範囲は非常に広い．しかしながら，このような顧客やユーザの行動履歴の情報から「商品間の相関関係をグラフ構造で表したい」，「映画間の評価の相関関係をグラフ構造で表したい」といった要請がある場合には，これらの関係性の有無が物理的に一意に定義されないため，分析者の観点に基づく適切なグラフ構造化が必要となってくる．たとえば「商品間の関係」の場合，あるアイテム A とアイテム B をノードとみなしたとき，顧客の購買履歴データに基づいて，これらの間にリンクを引くか否かを決める方法は一意ではない．単純な方法をいくつか考えてみても，

- アイテム A とアイテム B を同時刻（同じレシート上で）に買った顧客が 1 人でもいたらリンクを引く
- アイテム A とアイテム B を買った経験のある顧客が 1 人でもいたらリンクを引く
- アイテム A とアイテム B を買った経験のある顧客が 10 人以上いたらリンクを引く
- アイテム A を買った顧客の中で，ほかのアイテムも購入した顧客の数をそれぞれカウントしたとき，アイテム B も購入している顧客の数が最も多かっ

ごとう　まさゆき
早稲田大学創造理工学部経営システム工学科
〒 169–8555　東京都新宿区大久保 3–4–1
goto@it.mgmt.waseda.ac.jp

たらリンクを引く
・アイテム A を買った顧客の中で，ほかのアイテムも購入した顧客の数をそれぞれカウントしたとき，アイテム B も購入している顧客の数が上位 k 番目に入っていたらリンクを引く

など，さまざまな方法が考えられる．このような対象問題を一般化すると「ノード間に“ある種の類似度”が定義できる場合に，どのノード間にリンクを引き，どのノード間でリンクを引かずにグラフ構造化するか」という問題が見えてくる．単に，リンクの有無だけでなく，リンクに関係性の大きさを表す重みを付与した重み付きネットワークを構成する方法もあるが，類似度が完全に 0 となるノード間でのみリンクが切れることになるため，計算コストが膨大になることに加え，しばしばネットワーク構造が密になり，可視化の際に問題が生じることも多い．そのため，類似度の高いノード間でのみ，リンクを引いたような疎なネットワークを構築し，一般的なグラフマイニング手法を適用する方法が簡便で，かつ有用である．すなわち，いったん，比較的リンクの少ないグラフ構造ができると，既存のグラフマイニング手法を適用することで，さまざまな結果を得ることができる．一方で，その分析結果は，適用するグラフマイニング手法の以前に，グラフ構造の構築方法に強く依存してしまうことに注意が必要であろう．

本稿では，以上のような“単純にノード間の関係性の有無が定義できない問題”に対して，グラフマイニング手法を適用するケースを考え，一般的な問題の記述を行うとともに，k-NN (k-nearest neighbor) グラフと ε-NN (ε-nearest neighbor) グラフというモデル [4, 5] の応用について解説する．さらに，このような問題として扱うことができる対象問題について考察するとともに，ユーザの閲覧履歴データに基づいて Web サイト間のネットワーク分析を行った事例を紹介する．

2. 問題設定

本稿では，“単純にノード間の関係性の有無が定義できない問題”に対して，ネットワーク分析手法を適用するケースを対象とする．このような場合においても，ネットワーク分析やグラフマイニング手法を適用するためには，何らかの形でノード間の類似性が定義されている必要がある．ここでは，このような問題を一般的な形式で記述してみることにしよう．

2.1 グラフ構築手法

いま，ノード v_i とノード v_j の間の類似度を S_{ij} とする．このとき，ノード間の類似性が全くない状態，すなわち，$S_{ij} = 0$ であれば，ノード v_i とノード v_j の間にはリンクを引かず，$S_{ij} > 0$ であれば ノード v_i とノード v_j の間にはリンクを引くことを考える．このとき，多くの現実問題では，ほとんどのノード間で多少の類似性があるため，必要以上に多くのノード間でリンクが引かれてしまい，ネットワーク構造を可視化することのメリットが失われてしまう．

通常，SNS サイトにおける友人関係や Web サイト間のリンク関係のようなグラフ構造データでは，全体のノード数に比べて，実際にリンクの引かれる箇所は非常に少ないことがほとんどである．すなわち，グラフマイニングやネットワーク分析では，このような疎なグラフ構造データを想定しているといっても過言ではない．もし，すべてのノード間でリンクがある場合には，リンクの有無によるクラスタリングやラベル伝播手法は意味をもたなくなり，データ間の類似度行列を出発点とした分析手法の方が適切と考えられる．

一方で，実問題におけるノード間の類似度は，一部の少数のノード間で高い値を取り，そのほかの多くのノード間では非常に小さい値を取ることが多い．たとえば，商品間の類似性を“共通に購入した顧客数”で定義した場合，一緒に購入される頻度が高いアイテムの組み合わせは非常に少数である．このような場合，類似度の高いノード間でのみリンクを張ることで，疎なネットワークを構成することが可能である．このようなグラフ構造の構築法としては，以下のような方法が知られている [4].

1. **k-NN** グラフ： 各ノード v_i に対し，ほかのノードとの類似度 S_{ij} が高い上位 k 件のノードを選択し，それらとの間にリンクを引いたネットワークを構成する．
2. **ε-NN** グラフ： 類似度 S_{ij} が，$S_{ij} \geq \varepsilon$ となるようなノード v_i とノード v_j の間でリンクを引いたネットワークを構成する．

これらの手法において，リンクを引くノード数を決める k や類似度閾値である ε は分析者が適切に決める必要のあるパラメータである．

2.2 ノード間の類似度計算

たとえば，
・共通に購入した顧客数に応じて，商品間の類似性を定義したい
・共通に閲覧したユーザ数に応じて，Web サイト間の類似性を定義したい
・共通に高評価した評価者数に応じて，映画コンテ

ンツ間の類似性を定義したい

・共通にファンクラブに所属しているファン数に応じて，著名タレント間の類似性を定義したい

といった問題は，数学的には同じモデルで記述することができる．

ここでは，Web サイト間の類似性を定義する例で説明しよう．ある Web サイト v_i に対し，ユーザ u_l が閲覧していたら $a_{il}=1$，閲覧していなかったら $a_{il}=0$ という値を要素としてもつベクトル

$$\boldsymbol{v}_i = (a_{i1}, a_{i2}, \ldots, a_{iU})^\top \tag{1}$$

を定義する．ここで，U はユーザ数，$\top$ は転置を表す．このとき，Web サイト間の類似度 S_{ij} を，$\boldsymbol{v}_i$ と $\boldsymbol{v}_j$ の内積

$$S_{ij} = \boldsymbol{v}_i^\top \boldsymbol{v}_j \tag{2}$$

や余弦

$$S_{ij} = \frac{\boldsymbol{v}_i^\top \boldsymbol{v}_j}{\|\boldsymbol{v}_i\|\|\boldsymbol{v}_j\|} \tag{3}$$

により定義することが可能である．類似度尺度としては最大値が 1 となる余弦尺度 (3) のほうが類似度と呼ぶには相応しいが，一方で式 (2) の内積は“共通閲覧ユーザ数”という物理的な意味を有するため，どちらを採用するかは，分析目的やデータ特性を考慮して決めるべきであろう．

一方で，一般に各ユーザが閲覧する Web サイトは，全 Web サイトからすれば比較的少数であることから，これらのベクトル $\boldsymbol{v}_i$, $\boldsymbol{v}_j$ は 0 の要素が多く，1 が比較的少ないスパースな高次元ベクトルとなっている．このような高次元スパースなベクトル同士の内積も，それなりに意味をもつことは経験的に確かめられているが，インターネット利用時間が極端に長く，多数の Web サイトを閲覧しているユーザ，逆に極端に閲覧 Web サイト数の少ないユーザの影響を受けてしまう．また，類似した嗜好をもつ潜在的なユーザグループや内容が類似した Web サイトの潜在カテゴリが存在し，これらによって閲覧行動が全く異なっていることが考えられる．このようなデータに対し，高次元スパースなままのベクトル間で類似度を測ると，ノイズの影響を受けやすくなったり，潜在的なグループのサイズが大きいものの影響を強く受けてしまう．そこで，Web サイトとユーザを適切にクラスタリングする機能をもつ非負値行列因子分解 (NMF: Non-negative Matrix Factorization) [6] などの手法を用いて，次元圧縮を行うことが考えられる．いま，Web サイト—ユーザの閲覧／非閲覧情報を 0–1 で表現した $W \times U$ 行列

$$\boldsymbol{X} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1U} \\ a_{21} & a_{22} & \cdots & a_{2U} \\ \vdots & \vdots & \vdots & \vdots \\ a_{W1} & a_{W2} & \cdots & a_{WU} \end{pmatrix} \tag{4}$$

を考える．W は Web サイト数である．このとき，NMF は，$W \times K$ 行列 $\boldsymbol{N}$ と $K \times U$ 行列 $\boldsymbol{M}$ を用いて，

$$\boldsymbol{X} \approx \boldsymbol{N}\boldsymbol{M} \tag{5}$$

のように近似する手法である．具体的には，$\|\boldsymbol{X} - \boldsymbol{N}\boldsymbol{M}\|_F^2$ を最小化するように，$\boldsymbol{N}$ と $\boldsymbol{M}$ を逐次更新させて収束させる．ただし $\|\cdot\|_F$ はフロベニウスノルムである．このように行列分解をすると，$W \times K$ 行列 $\boldsymbol{N}$ の各行ベクトルは，各 Web サイトを K 個の潜在変数への所属度を用いて表現したものとなっている．すなわち，元の U 次元のスパースベクトルから，K 次元のベクトルへと低次元化することができる．すなわち，

$$\boldsymbol{N} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1K} \\ b_{21} & b_{22} & \cdots & b_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ b_{W1} & b_{W2} & \cdots & b_{WK} \end{pmatrix} \tag{6}$$

と記述し，各 Web サイトベクトルを，K 次元ベクトル

$$\tilde{\boldsymbol{v}}_i = (b_{i1}, b_{i2}, \ldots, b_{iK})^\top \tag{7}$$

で再定義すれば，これらの内積，もしくは余弦によって，Web サイト間の類似性 S_{ij} を定義することが可能である．すなわち，

$$S_{ij} = \frac{\tilde{\boldsymbol{v}}_i^\top \tilde{\boldsymbol{v}}_j}{\|\tilde{\boldsymbol{v}}_i\|\|\tilde{\boldsymbol{v}}_j\|} \tag{8}$$

とすることで，ノイズが除去され，ユーザや Web サイトの潜在的な類似グループの統計的傾向を反映した類似度が構成できる．

もとのベクトル間の類似度式 (2), (3) を用いるか，あるいは NMF によって次元圧縮してから得られる類似度式 (8) を計算するかは，分析者の判断による．一般には，複数の嗜好の異なるユーザが混在しており，Web サイトのほうもさまざまな特徴の異なるサイトが混在していることが想定できる場合には，NMF による次元圧縮を行い，各潜在変数への所属度を用いて類

似度を計算するほうがよいと考えられる．

3. ユーザの閲覧履歴データに基づく Web サイト間のネットワーク分析

ここでは，株式会社ヴァリューズより提供いただいたインターネット上の Web サイト閲覧履歴データを用いて，Web サイト間の関係性について，クラスタリング分析によって可視化した結果を紹介する．本データに対して Web サイトの関係性を分析する手法としては，他にもさまざまな方法が考えられ，最近注目を集めている分散表現モデルによる方法も有用である．このような手法の分析結果については，文献 [7] などを参照されたい．

このデータは，登録に同意したモニタが PC またはスマートフォンから閲覧した Web サイトのホスト名を記録したデータであり．データ取得期間は，2017 年 8 月 1 日から 2017 年 10 月 31 日の 3 カ月である．総閲覧数は 7,224,737 回，ユーザ数は 9,851 人，Web サイト数は 27,789 である．しかし，これらのすべてを分析対象とすると，Google や Yahoo!などの閲覧回数が桁違いに多いサイトの影響で興味深い結果が得られにくい．また，閲覧回数が非常に少ない多数の Web サイトが存在するため，これらの多数の Web サイトをノードとして取り込んだネットワーク分析も，大多数の孤立ノードを生むだけの結果を導いてしまうという問題がある．そこで，閲覧回数が 100 回以上 2,000 回未満の Web サイトを抽出し，かつ「クーポン・ポイント」と「アダルト」のカテゴリ情報が付与された Web サイトを除外して分析対象とした．その結果，対象となるユーザ数は 991 名，Web サイト数は 4,965 となった．

また，分析環境は Jupyter Notebook であり，グラフ可視化ライブラリとして Networkx [8] を用い，ノードとリンクの座標情報を与えて可視化している．ノードの座標計算手法には，spring_layout 関数 [9] を用いており，以下の図で使われている座標系はこれによるものである．これらの元となっているグラフ描画アルゴリズムについては，文献 [10] を参照されたい．

3.1 k-NN グラフを用いたネットワーク分析

まず，単純に“共に閲覧したユーザ数”を Web サイト間の類似度とした k-NN グラフにおいて，$k = 1, 2, 5, 10$ と変化させた場合のネットワーク可視化の結果を，図 1〜図 4 に示す．

$k = 1$ は，各 Web サイトについて，最もともに閲覧しているユーザの数が多い Web サイト一つとのみリンクが繋がっているため，ネットワークがかなりスパースである（図 1）．これに対し，$k = 2$ とすると，一つの Web サイトから類似性の高い二つの Web サイトとのリンクが張られるが，この段階ですでに密なネットワークとなっており，クラスタに分かれていない（図 2）．さらに，$k = 5, 10$ と大きくすると，グラ

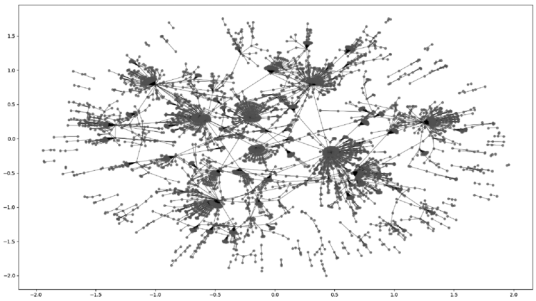

図 1　k-NN グラフによる可視化結果（$k = 1$ の場合）

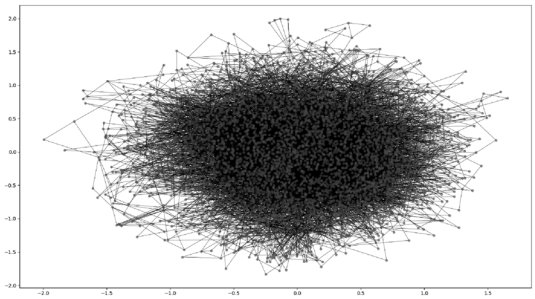

図 2　k-NN グラフによる可視化結果（$k = 2$ の場合）

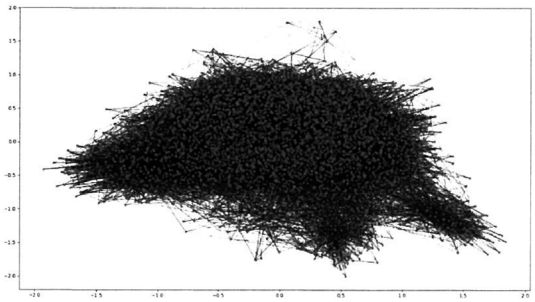

図 3　k-NN グラフによる可視化結果（$k = 5$ の場合）

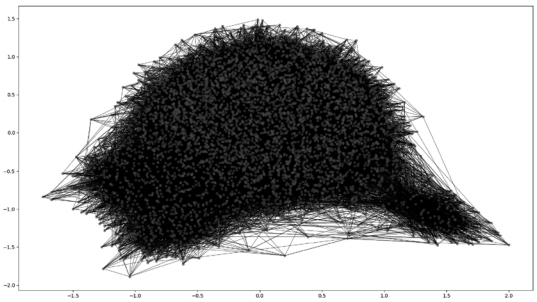

図 4　k-NN グラフによる可視化結果（$k = 10$ の場合）

フ構造からクラスタリングをしても，ほとんどの Web サイトが中央で一塊になってしまい，あまり興味深い結果が得られていないことがわかる（図 3，図 4）．

以上を総合すると，$k = 1$ のケースでは，多少のコミュニティが存在することが見て取れる結果であるが，$k \geq 2$ とするとほとんどのノードが相互に結合してしまい，コミュニティらしきものを発見することができていないことがわかった．k-NN グラフを用いたクラスタリングでは，ほかの多くのノードと関係を有するノードであっても，類似度の高い上位 k 件のノードとのリンクを残してほかは切断されてしまうことから，全体的に均質な接続関係になってしまっている可能性がある．

3.2 ε-NN グラフを用いたネットワーク分析

前節の k-NN グラフでは有用なコミュニティを発見できなかったが，ε-NN グラフでは，類似度が高いノード間のすべてでリンクが張られることから，ほかの多くのノードとのリンクを有した，重要度の高いノードが生成されやすいことが期待できる．“共に閲覧したユーザ数”を Web サイト間の類似度とし，ε-NN グラフにおいて，ε を変化させた場合のネットワーク可視化の結果を，図 5〜図 8 に示す．

$\varepsilon = 1$ の場合は，“共に閲覧したユーザ数”が 1 人でも存在した場合に，Web サイト間でリンクが貼られるため，多くの Web サイト間で密な結合をもったネットワークになる．そのため，ネットワーク可視化の結果も，図 5 のように，ほとんどのノードが中心で一塊となり，周辺にリンク数の少ないノードが散見されるような構造になる．ε を $\varepsilon = 5, 20, 50$ と大きくするとともにリンク数は少なくなるが，中央で多くのノードが結合し，周辺で孤立したノードが存在する傾向は変わらない．

以上のように，単純に“ともに閲覧したユーザ数”を Web サイト間の類似度として，k-NN グラフや ε-NN グラフを生成して，クラスタリング分析をしても，あまり特徴的なコミュニティが現れないことがわかる．また，k-NN グラフにおける k や ε-NN グラフにおける ε といったパラメータの設定によって結果も変わるため，これらのセッティングも分析者にとっては難しい作業となってしまう．

3.3 NMF 類似度による ε-NN グラフを用いたネットワーク分析

前節で示したように，高次元スパースなベクトル間の内積を類似度として，k-NN グラフや ε-NN グラフを生成して，クラスタリング分析をしても，明確なコミュニティを得ることは難しかった．そこで，NMF を用いて次元を圧縮し，潜在クラスへの帰属度ベクトルを用いて，各 Web サイト間の類似度を定義し，ε-NN グラフを生成してみる．この NMF-ε-NN グラフに対し，ネットワーク構造を可視化した結果を図 9 に示す．

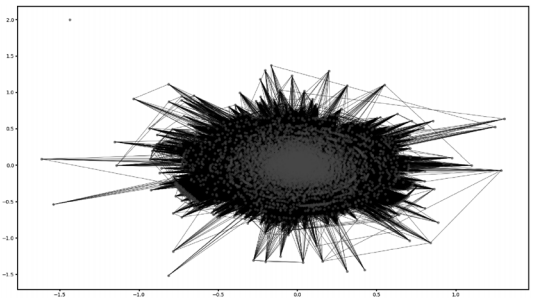

図 5　ε-NN グラフによる可視化結果（$\varepsilon = 1$ の場合）

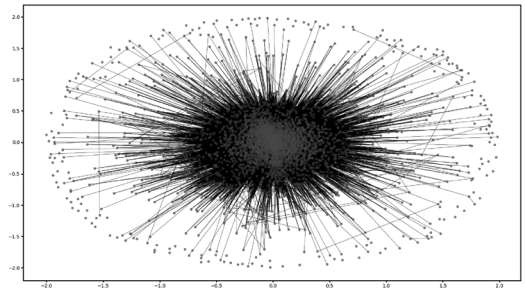

図 6　ε-NN グラフによる可視化結果（$\varepsilon = 5$ の場合）

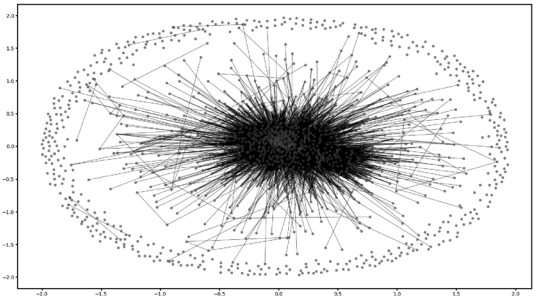

図 7　ε-NN グラフによる可視化結果（$\varepsilon = 20$ の場合）

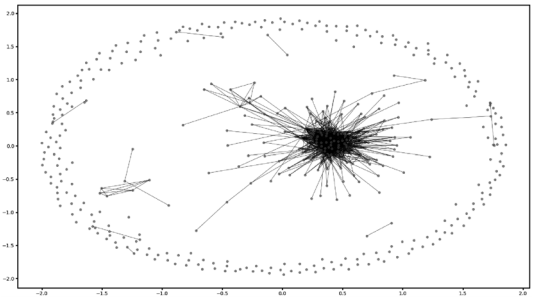

図 8　ε-NN グラフによる可視化結果（$\varepsilon = 50$ の場合）

ただし，この例では，リンクを張る類似度の閾値として $\varepsilon = 0.999$ と設定した．すなわち，類似度が 0.999 以上の極めて類似性の高い Web サイト間でリンクが張られたグラフ構造に対し，クラスタリングを適用した結果が図 9 である．この結果より，いくつかの密に接続しているコミュニティ（ノード集合）が見られるようになり，これらがクラスタを形成していることがわかる．

図 9 ではコミュニティが形成されるようになったので，これらの中身を確認し，クラスタを形成している Web サイト群の内容を追記したものが図 10 である．

その結果，図 10 に示したように，コミュニティを形成している Web サイト群に，その特徴を付与することができている．これらのクラスタは，Web サイト間のハイパーリンク構造から抽出されたコミュニティではなく，あくまで「ユーザがともに閲覧するか否か」という閲覧履歴データから得られているネットワーク構造であることに注意したい．したがって，互いにハイパーリンクが張られていない Web サイト間であっても，同じ嗜好をもったユーザから閲覧されている回数が多ければ，このグラフ構造においてはリンクが張られていることになる．

以上のように，閲覧するユーザの共起性という観点から，Web サイト間の類似性でクラスタリング分析する場合，単に“ともに閲覧したユーザ数”を Web サイト間の類似度とした k-NN グラフや ε-NN グラフではなく，NMF を用いて次元縮約してから類似性を計算した場合の ε-NN グラフによって，興味深いクラスタリング結果が得られることがわかる．このように，ユーザや顧客の行動履歴データを用いてネットワーク分析を行う場合には，パラメータの設定だけでなく，データの次元削減などの操作も必要となり，このようなさまざまな試行錯誤は分析者のスキルに依存すると言えよう．

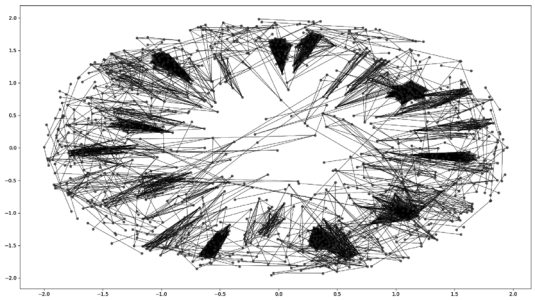

図 9　NMF 類似度を用いた ε-NN グラフによる可視化結果

4.　その他の応用事例

前節では，ユーザの閲覧履歴データに基づいて Web サイト間の類似性を定義し，得られたグラフ構造に対してクラスタリング分析を行った．グラフ構造が作られれば，重要なノードを特定する“重要度分析”や一部のノードに付与されたラベルから，周囲のノードのラベルを推定する“ラベル伝播”などの分析手法も適用することができる．

類似度からグラフ構造を構成し，これらのグラフマイニング手法が有用となりうる実問題としては，たと

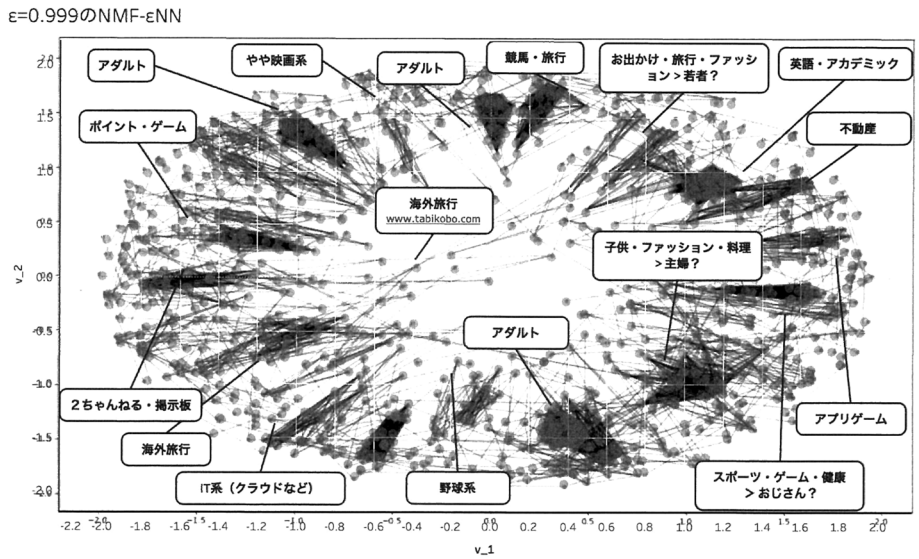

図 10　NMF 類似度を用いた ε-NN グラフによるコミュニティ抽出

えば，下記のような例が挙げられる．

- 共通に購入した顧客数に応じて，商品間の類似性を定義し，クラスタリングによって類似した商品群を抽出することで，商品戦略に結び付ける．加えて，重要度分析によって，多くの商品とあわせて購入されるような重要商品を特定することもできる．
- 共通に高評価した評価者数に応じて，映画コンテンツ間の類似性を定義し，グラフ構造を得ることで，あるユーザの評価値履歴データから，ラベル伝播によって，そのほかのコンテンツの評価値を予測する．予測された評価値に基づき，各ユーザに高評価が予測されるコンテンツを推薦することができる．
- コインパーキングをノードとし，それらの間の距離を用いて，グラフ構造を得ることで，一部のコインパーキングの利用状況（満車か否かのラベル）から，ラベル伝播によって，そのほかのコインパーキングの利用状況を推定する．リアルタイムに駐車可否を知らせるシステムにおいて，一部のデータを取得するだけで，広範囲なコインパーキングの状況推定が可能となる．
- 就職ポータルサイト上で，同じ学生ユーザからのエントリーの共起によって，採用活動中の企業間の類似度を定義し，クラスタリングによって類似企業群を抽出する．これにより，就職活動中の学生への適切な情報推薦や効率的な情報検索の仕組みを構築できる可能性がある．

以上のうち，顧客の購買履歴に基づく商品間の類似性からグラフマイニングを適用した例としては，無印良品ブランドに対してネットワーク分析を適用した伊藤らの研究 [11] がある．この事例では，約 1 年間の ID-POS データを活用し，商品をともに購入した顧客数によって，商品間の類似性を定義し，ネットワーク分析を用いて，商品をクラスタリングして購買傾向という観点から類似性の高い商品グループを特定している．加えて，年間購買金額によって顧客を層別し，より購買金額の高い優良顧客にとって重要な商品を分析している．このような分析は，各顧客の優良顧客への成長を促す戦略を立案する際に有用な情報となりうるであろう．

5. おわりに

本稿では，顧客やユーザの行動履歴の情報を用いて，グラフ構造を構築し，ネットワーク分析を行う方法について解説を行った．単純に，高次元スパースなノード間の類似度を計算し，k-NN グラフや ε-NN グラフを生成した場合には，興味深い分析結果が得られなかったり，パラメータの調整が難しいことも多い．一方で，NMF などの適切な次元圧縮手法と併用して，グラフ構造を作成すると，関係が深いノード同士が結合し，興味深いクラスタリング結果が得られる場合があることを示した．

分析対象の関係性についてある種の類似度が定義でき，かつ類似度が高いものは比較的少数で，ほとんどの類似度が低い場合，k-NN グラフや ε-NN グラフによって疎なネットワークを構成することができ，そのグラフ構造にクラスタリングや重要度分析，ラベル伝播などの分析手法をそのまま適用することが可能となる．疎なグラフ構造を統計的モデリングの枠組みで構成するためには，正則化手法を駆使したスパースモデリングなども有用となりうるが [12]，本稿で示した k-NN グラフや ε-NN グラフであっても，NMF などの適切な次元圧縮手法と組み合わせれば，比較的容易に有用な情報を得たり，ネットワーク構造を可視化したりすることができる．今後も，このような比較的簡便なネットワーク分析の適用事例が増え，さまざまな分野で活用されていくことを期待したい．

謝辞　本稿の執筆にあたり，貴重なデータを提供頂いた株式会社ヴァリューズに感謝します．また，実データの整形と分析をサポートしてくれた早稲田大学・後藤研究室の保坂大樹君に感謝します．

参考文献

[1] 飯田恭弘, 岸本康成, 藤原靖宏, 塩川浩昭, 鬼塚真, “大規模グラフ構造データからのコミュニティ抽出と重要度計算 —高速化への取組みと応用—,” 人工知能, **29**, pp. 472–479, 2014.
[2] 鬼塚真, “グラフマイニング技術を用いたビッグデータの応用と高速化技術の取り組み,” 生産と技術, **67**(2), pp. 4–10, 2015.
[3] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” In *Proceedings of the National Academy of Science*, **99**, pp. 7821–7826, 2002.
[4] L. N. Ferreira and L. Zhao, “A time series clustering technique based on community detection in networks,” *Procedia Computer Science*, **53**, pp. 183–190, 2015.
[5] A. Clauset, M. E. J. Newman and C. Moore, “Finding community structure in very large networks,” *Physical Review E*, **70**, article number: 066111, 2004.
[6] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, **401**, pp. 788–791, 1999.
[7] 保坂大樹, 河部瞭太, 山下遥, 後藤正幸, “意味空間上の

分布表現に基づく Web サイトと閲覧ユーザの統合分析モデル，”情報処理学会論文誌，**60**, pp. 1390–1402, 2019.
[8] https://networkx.github.io/documentation/stable/index.html（2019 年 8 月 20 日閲覧）
[9] https://networkx.github.io/documentation/networkx-1.9/reference/generated/networkx.drawing.layout.spring_layout.html（2019 年 8 月 20 日閲覧）
[10] T. M. J. Fruchterman and E. M. Reingold, “Graph drawing by force-directed placement,” *Software: Practice and Experience*, **21**, pp. 1129–1164, 1991.
[11] H. Ito, G. Kumoi and M. Goto, “A study on extraction of important items focused on customer growth based on network analysis,” In *Proceedings of the 18th Asia Pacific Industrial Engineering and Management System Conference*, ID164, 2017.
[12] R. Kawabe, H. Yamashita and M. Goto, “An analysis of web access log data based on graph mining method,” In *Proceedings of the 19th Asia Pacific Industrial Engineering and Management System Conference*, ID153, 2018.