距離減衰重みを導入したノード群へのアノテーション付与法

伏見 卓恭[1,2,a)]　佐藤 哲司[1]　斉藤 和巳[3]　風間 一洋[4]

概要：ネットワークにおいて，ノード群が密に隣接する部分をコミュニティとして抽出することは，大規模なネットワークの構造理解や現象分析の点で非常に重要である．しかし，抽出したコミュニティにおいて，属するノード同士が必ずしも類似の特徴を有するとは限らない．本研究では，各ノードの活動やコンテンツから得られる特徴ベクトルに，距離減衰重みを導入した合成ベクトルを考える．それらをクラスタリングすることで意味的なまとまりのあるノード群を抽出し，各ノード群に有意に出現する特徴量でアノテーションする方法を提案する．また，全てのノードが近隣ノードと同程度の類似傾向にあるわけではなく，ノードによって減衰の程度を調整しなければならない．従って，減衰重みを制御する指数減衰関数のパラメータをノードごとに推定する方法についても説明する．複数の実データを用いて，推定したパラメータとノードの性質について考察し，抽出したノード群へのアノテーションについても評価する．

1. はじめに

TwitterやFacebook，CookpadなどのSNSや，レビューサイト，ブログサイトなどのソーシャルメディアにおいて，ユーザ間に多くのインタラクションが存在する．それらをネットワークとしてとらえ分析することにより，様々な知見が得られている．このようなネットワークにおいて隣接関係にあるユーザ間には，共通の特徴があると考えられる [1]．例えば，化粧品に関するレビューサイトを利用するユーザを，商品に対するレビュー評点を要素とする商品次元のベクトルで表現する．この時，隣接関係にあるユーザ間のベクトルは比較的類似する傾向にある．しかし，ネットワーク上のすべてのユーザ同士が類似ベクトルを有するとは，通常考えられない．すなわち，ネットワークのどこかで嗜好の切れ目が存在すると考えられる．さらに，各ノードは隣接関係にあるノードと類似特徴を有する傾向にはあるが，その傾向はノードによってさまざまである．

本研究では，特徴ベクトルの類似性に着目した意味的まとまりと，ノード間の隣接関係に着目した構造的まとまりの両方を考慮したノード群を抽出し，そのノード群にアノテーションする手法を提案する．意味的まとまりのノード群を抽出するなら，特徴ベクトルを単純にクラスタリングすることで実現できる．しかし，これらはノードの連結性を保証していないため，ノード群（部分グラフ）へアノテーションする本稿の目的には不十分である．逆に，構造的まとまりのノード群を抽出するなら，CNM法に代表されるコミュニティ抽出法で実現できるが，CNMコミュニティ内に複数の意味的まとまりがある場合には不十分である．提案手法では，各ノードの特徴ベクトルに対し，ノード間距離に基づく減衰重みを導入した合成ベクトルを考える．減衰重みを付した合成ベクトルであるため，特徴ベクトルとしての意味は幾分か均されるが，隣接関係にあるノード群は類似の合成ベクトルを得やすくなる．この合成ベクトルをクラスタリングすることで，意味的なまとまりの強い連結ノード群を抽出する．そして，抽出したノード群に対し，そのノード群に有意に多く出現する特徴量をラベルとして付与することにより，どのようなノード群なのかをアノテートする．提案手法により，共通あるいは類似の特徴量を有するノード群は，ネットワークのどのあたりに位置するかなどを把握することが出来るようになる．

新たに導入した距離減衰重み付きの合成ベクトルであるが，隣接するノード群は類似の特徴ベクトルを有することを前提としている．しかし，すべてのノードが同程度に隣接ノードと類似するとは限らない．そこで，各ノードに対

[1] 筑波大学　図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba
[2] 日本学術振興会特別研究員 (PD)
JSPS research fellow (PD)
[3] 静岡県立大学　経営情報学部
School of Management and Information, University of Shizuoka
[4] 和歌山大学　システム工学部
Faculty of Systems Engineering, Wakayama University
[a)] takayasu.fushimi@gmail.com

して適切な減衰パラメータを推定する．

2. 関連研究

意味的特徴と構造的な特徴の両方を用いたコミュニティ抽出手法として，Kuramochi らの手法がある [2]．この手法では，与えられたグラフ構造から，極大クリークなどの密なノード集合をノード，クリーク間のリンクをリンクとした交グラフを構築する．交グラフにおけるノード間のリンクには，特徴量より算出する重みを付与する．この際に，交グラフのノード（密なノード集合に相当）内のノードの特徴量を併合し，TF・IDF をかけている．本研究の提案手法でも，周辺ノードの特徴ベクトルを合成するが，距離に従って減衰させながら合成する点，および，減衰の強弱をノードごとに推定する点で異なる．

Wu らは，与えられたネットワークに対し，ノード間の類似度などを重みとした Conceptual ネットワークにおける重みの和が最大で，Physical ネットワーク（実際の接続関係）において連結となる Densest Connected Subgraph を抽出する手法を提案している [3]．この手法では，低次数ノードを枝刈りすることで効率的なアルゴリズムを実現しているが，本研究では，与えられた全てのノードを構造的・意味的つながりのあるノード群に分割する点で異なる．いわば，コア抽出とコミュニティ分割の違いに相当する．

アノテーション法の関連研究として，小林らの可視化結果へのアノテーション法がある [4]．この手法では，2 次元上に可視化されたオブジェクト群に対して，可視化座標の近接性に基づいて最小全域木を構築する．そして，オブジェクト群が有する特徴量に基づく尤度関数を定義し，尤度関数が最大になるリンクを切断することにより，オブジェクト群を部分集合に分割していく．最終的に得られた部分集合群は，特徴的な特徴量分布を有する部分集合になっている．各部分集合に対し，Z スコアにより統計的有意に出現する特徴量を抽出し，アノテーションのラベルとして採用する．本研究では，ネットワーク構造を対象としている点で，最小全域木を切断する小林らの手法と関連する．しかし，ネットワーク構造は木構造と異なり，一般に 1 本のリンクを切断することによりノード集合を分割することができない．したがって，小林らの手法をそのまま適用することはできない．

また，ナイーブな手法として，各ノードの特徴ベクトルをそのままクラスタリングする方法が考えられるが，それでは，ノード間の隣接関係を考慮できない．逆に，Clauset らの CNM クラスタリング [5] により，ノード群をコミュニティに分割する方法も考えられるが，これでは，コミュニティ内の異なる特徴の共存に対応できない．

3. 提案手法

提案手法は，ネットワーク構造 $G = (V, E)$，各ノード $u \in V$ の J 次元特徴ベクトル $\mathbf{x}_u$ およびクラスタ数 K を入力とし，以下の手順でノード群を抽出し，アノテーションを付与する．

CA1 各ノードを中心に，距離減衰合成ベクトルを構築する；

CA2 距離減衰合成ベクトルを K-medoids 法によりクラスタリングする；

CA3 各クラスタに統計的有意に出現する特徴量を Z スコアにより算出する；

CA4 Z スコア上位の特徴量により，各クラスタにアノテートする；

以下では各手順を詳しく説明する．

3.1 距離減衰合成ベクトル

各ノード u に対して，他ノード v へのグラフ距離（最短パス長）を $d(u,v)$ とする．ただし，$d(u,v) = d(v,u)$ であり，$d(u,u) = 0$ である．各ノード u を中心に，隣接するノード v の特徴ベクトル $\mathbf{x}_v$ を，指数的減衰関数 $\exp(\lambda d)$ を用いて距離 d に従って減衰させながら合成ベクトルを構築する；

$$\begin{aligned}\mathbf{y}_u &= \sum_{d=0}^{D_u} \exp(-\lambda d) \sum_{v \in \Gamma_d(u)} \mathbf{x}_v \\ &= \sum_{v \in V} \exp(-\lambda d(u,v)) \mathbf{x}_v. \end{aligned} \quad (1)$$

ここで，$D_u = \max_{v \in V}\{d(u,v)\}$ はノード u から全ノードへ到達するための最大グラフ距離を表し，$\Gamma_d(u)$ はノード u から距離 d にあるノード集合を表し，$\Gamma_0(u) = \{u\}$ とする．このベクトル $\mathbf{y}_u$ を，ノード u を中心とした距離減衰合成ベクトルと呼ぶ．距離減衰合成ベクトルは，ノード u と直接隣接，あるいは，近くに存在するノードの特徴ベクトルを重みを強くして足し込む．逆に遠くに存在するノードの特徴ベクトルを重みを弱くして足し込んでいる．従って，ベクトルとしては幾分か均されているが，隣接するノード間で類似のベクトルになりやすくなる．さらに，遠くに類似の特徴ベクトルを有するノードが存在しても，周辺ノードの特徴ベクトルが異なる場合は，異なる距離減衰合成ベクトルになるため，クラスタリング結果として，異なるクラスタになる．

3.2 K-medoids クラスタリング

距離減衰合成ベクトルに対して K-medoids 法 [6] を用いて，全ノード集合 V を K 個のノード群 $\{V_1, V_2, \ldots, V_K\}$ に分割する．

K-medoids 法は，オブジェクト集合 V とその要素 $v, w \in V$ 間の類似度 $\rho(v,w)$ が与えられたとき，以下の目的関数を最大にするような代表オブジェクト集合 P を求める．

$$\mathcal{J}(P) = \sum_{v \in V} \max_{w \in P} \{\rho(v, w)\}.$$

$K = |P|$ 個の代表オブジェクトを抽出し，残りのオブジェクト群を最も類似する代表オブジェクトのクラスタに割り当てることで，オブジェクト集合を K 個のクラスタに分割する．K-medoids 法の解法には反復法や貪欲法があるが，K-means 法と異なり解の一意性が保証される貪欲法を用いる．本稿では，類似度 $\rho(u, v)$ は，距離減衰合成ベクトル間のコサイン類似度とする：

$$\rho(u, v) = \frac{\mathbf{y}_u^T \mathbf{y}_v}{||\mathbf{y}_u|| ||\mathbf{y}_v||}.$$

3.3 Z スコア

各ノード群 V_k に有意に多く出現する特徴量を Z スコアを用いて抽出する．ここで，ネットワーク全体の特徴量の分布を以下のように定義する：

$$p_j = \frac{\sum_{u \in V} x_{u,j}}{M}.$$

ここで分母の M は，確率にするための正規化項であり，$M = \sum_{v \in V} \sum_{j=1}^{J} x_{v,j}$ である．また，ノード群 V_k に属するノードの特徴量を以下のように合算する：

$$q_j^{(k)} = \sum_{u \in V_k} x_{u,j}.$$

この時，ノード群 V_k に対する特徴量 j の Z スコアは以下のように計算する：

$$z_j^{(k)} = \frac{q_j^{(k)} - M_k p_j}{\sqrt{M_k p_j (1 - p_j)}}.$$

ここで，$M_k = \sum_{v \in V_k} \sum_{j=1}^{J} x_{v,j}$ である．ノード群 V_k に特徴量 j が出現する期待値 $(M_k p_j)$ に対して有意に多いか少ないかにより，各ノード群の特徴的な特徴量を抽出する．提案手法では，各ノード群ごとに Z スコア上位 H 件の特徴量をアノテーション用のラベルとして採用する．

4. パラメータ推定法

上述した距離減衰合成ベクトル $\mathbf{y}_u$ は，近隣ノードの特徴ベクトルを減衰させながら合成させて構築する．これは，近隣ノードの特徴ベクトルは類似する傾向にあるという前提がある．しかし，全てのノードが近隣ノードと同程度の類似傾向にあるわけではなく，ノードによって減衰の程度を調整しなければならない．本稿では，各ノードの距離減衰合成ベクトルが特徴ベクトルとコサイン類似度の意味で最も類似するように各ノードのパラメータ λ_u を設定する．ノルムを 1 に正規化した特徴ベクトルを $\mathbf{x}_u$ とし，

$$F_u(\lambda_u) = \mathbf{x}_u^T \frac{\sum_{v \in V \setminus \{u\}} \exp(-\lambda_u d(u, v)) \mathbf{x}_v}{||\sum_{v \in V \setminus \{u\}} \exp(-\lambda_u d(u, v)) \mathbf{x}_v||} \qquad (2)$$

という目的関数を考える．ここでは，距離減衰合成ベクトルの計算に自身の値を合成していない点で式 1 と異なることに注意されたい．

目的関数 2 を最大化するようなパラメータ λ_u を求める手順を説明する．ノード u に対して，距離 d にあるノードの特徴ベクトルの合成ベクトルを

$$\mathbf{f}_{u,d} = \sum_{v \in \Gamma_d(u)} \mathbf{x}_v$$

とし，ノード u の特徴ベクトルとの内積を

$$g_{u,d} = \mathbf{x}_u^T \mathbf{f}_{u,d}$$

とする．そして，ノード u からの距離の和が d になる合成ベクトル $\mathbf{f}_{u,d_1}$ と $\mathbf{f}_{u,d_2}$ ペア間の内積を足し合わせ

$$h_{u,d} = \sum_{d_1 + d_2 = d} \mathbf{f}_{u,d_1}^T \mathbf{f}_{u,d_2}$$

とすると，式 2 は以下のように書き換えられる：

$$F_u(\lambda_u) = \frac{\sum_{d=1}^{D_u} \exp(-\lambda_u d) g_{u,d}}{\sqrt{\sum_{d=2}^{2D_u} \exp(-\lambda_u d) h_{u,d}}}.$$

計算の便宜上，対数をとった以下の目的関数を最大にするようなパラメータ λ_u を求める：

$$\log F_u(\lambda_u) = \log \sum_{d=1}^{D_u} \exp(-\lambda_u d) g_{u,d} - \frac{1}{2} \log \sum_{d=2}^{2D_u} \exp(-\lambda_u d) h_{u,d}. \qquad (3)$$

ここで，事後確率関数を

$$r_{u,d} = \frac{\exp(\lambda_u d) g_{u,d}}{\sum_{d'=1}^{D_u} \exp(\lambda_u d') g_{u,d'}}$$

とすると，式 3 は以下のように書き換えられる：

$$\log F_u(\lambda_u) = \sum_{d=1} \bar{r}_{u,d} \{(-\lambda_u d) + \log g_{u,d}\} - \sum_{d=1} \bar{r}_{u,d} \log r_{u,d} - \frac{1}{2} \log \sum_{d=2}^{2D_u} \exp(-\lambda_u d) h_{u,d}.$$

パラメータ λ_u に関係のない項などを除くと

$$Q_u(\lambda_u) = -\lambda_u \sum_{d=1} \bar{r}_{u,d} \cdot d - \frac{1}{2} \log \sum_{d=2}^{2D_u} \exp(-\lambda_u d) h_{u,d}$$

となり，1 階微分は，

$$\frac{dQ_u(\lambda_u)}{d\lambda_u} = -\sum_{d=1} \bar{r}_{u,d} \cdot d + \frac{\sum_{d=2}^{2D_u} \exp(-\lambda_u d) \cdot d \cdot h_{u,d}}{2 \sum_{d=2}^{2D_u} \exp(-\lambda_u d) h_{u,d}}$$

となる．ここで，

$$s_{u,d} = \frac{\exp(-\lambda_u d) h_{u,d}}{\sum_{d'=2}^{2D_u} \exp(-\lambda_u d') h_{u,d'}}$$

とすると 2 階微分は，

$$\frac{d^2Q_u(\lambda_u)}{d\lambda_u^2} = -\frac{1}{2}\left\{\sum_{d=2}^{2D_u} s_{u,d}\cdot d^2 - \left(\sum_{d=2}^{2D_u} s_{u,d}\cdot d\right)^2\right\}$$

となり，ブレースの中は 2 次のモーメント同様に非負であるため，2 階微分自体は常に 0 以下となる．1 階微分が λ_u に関して閉じた形で書けないため，本研究ではニュートン法によりパラメータを求める．この推定されたパラメータは，値が大きいほど距離減衰の程度が強く，近隣ノードの値のみを大きな重みで，遠くのノードの値はほとんど無視する．逆に値が 0 に近いほど距離減衰の程度は弱く，近隣も遠方も同程度の重みで合成する．すなわち，近隣に類似の特徴ベクトルを有するノードが存在するか否かにより値が異なり，局所的なノード集合の中に順応しているノードは値が大きく，近隣に類似するノードが存在しない異端児ノードの場合は，多くのノードの特徴ベクトルを均等に合成しなければコサイン類似度を高くできないため，値が 0 に近くなる．

推定したパラメータによる距離減衰合成ベクトルを用いることで近隣に類似ノードが存在しないノードや近隣のみに類似ノードが存在するノードに関して，必要以上に特徴ベクトルを均すことを避けられるため，入力された特徴ベクトル $\mathbf{x}$ をより反映したクラスタリング結果が期待できる．

特徴ベクトルの次元を H，平均ノード間距離を $\bar{D}$ とすると，全ノードのパラメータ推定に要する時間計算量は，$h_{u,d}$ 計算に要する $O(|V|\times 2\bar{D}\times H)$ である．本提案手法では，様々な次元圧縮技術を用いて圧縮したベクトルを用いることも可能である．

5. 比較に用いる手法

この節では，提案手法と比較する手法について説明する．1 つ目は，特徴ベクトルをそのまま K-medoids 法によりクラスタリングする手法である．類似度は提案手法と同様，以下のようにコサイン類似度 $\rho(u,v) = \mathbf{x}_u^T\mathbf{x}_v/||\mathbf{x}_u||||\mathbf{x}_v||$ を用いる．各ノードの特徴ベクトル $\mathbf{x}$ をそのまま用いている点で，提案手法と異なることに注意されたい．この手法では，構造的なまとまりは一切考慮せず，意味的まとまりのみを対象としている．

2 つ目は，Clauset らによって提案された CNM 法である [5]．以下に示す，リンク構造に基づくモジュラリティ $Q = \sum_{k=1}^{K}\left(e_{kk} - a_k^2\right)$ を最大化するようにノードを分割し，コミュニティを抽出する．ここで，e_{kk} は，全リンク数に対するコミュニティ k 内のリンク数の比率を表し，$a_k = \sum_{h=1}^{K} e_{kh}$ は，コミュニティ k のノードが持つリンク数の比率を表している．この手法では，意味的なまとまりを陽に考慮せず，構造的なまとまりを対象としている．

3 つ目は，小林らによって提案された MST 分割法である [4]．この手法では，2 次元上に可視化されたオブジェクト群に対して，可視化座標の近接性に基づいて最小全域木を構築する．そして，オブジェクト群が有する特徴量に基づく以下の尤度関数を定義し，尤度関数が最大になるようにリンクを $K-1$ 本切断することにより，オブジェクト群を K 個の部分集合に分割していく．我々の提案手法における Z スコア計算の記述 $q_j^{(k)}$ を利用すると，$L = \sum_{k=1}^{K-1}\sum_{j=1}^{J} q_j^{(k)}\log q_j^{(k)}/q^{(k)}$ となる．ここで，$q^{(k)} = \sum_{j=1}^{J} q_j^{(k)}$ である．最終的に得られた部分集合群は，特徴的な特徴量分布を有する部分集合になっている．本稿では，隣接関係を反映した可視化結果を出力できるクロスエントロピー法 [7] により可視化する．

6. 評価実験

6.1 ネットワークデータ

1 つ目のネットワークは，とある大学のウェブサイトにおけるハイパーリンク構造である [*1]．ウェブページをノード，ハイパーリンクを無向化しリンクとした．各ノードの特徴ベクトルは，ウェブページの内容を形態素解析してえられる名詞群の Bag of Words とした．ノード数は 600，リンク数は 1,299，特徴ベクトルの次元数は 4,412 である．本稿では Web ネットワークと呼ぶ．

2 つ目のネットワークは，日本語ウィキペディア [*2] の人名の共起ネットワークである．人物記事をノード，5 つ以上の記事において共起関係のある人物間にリンクを張った無向ネットワークである．各ノードの特徴ベクトルは，記事内に出現する名詞群の Bag of Words とした．ノード数は 9,481，リンク数は 122,522，特徴ベクトルの次元数は 20,411 である．本稿では Wiki ネットワークと呼ぶ．

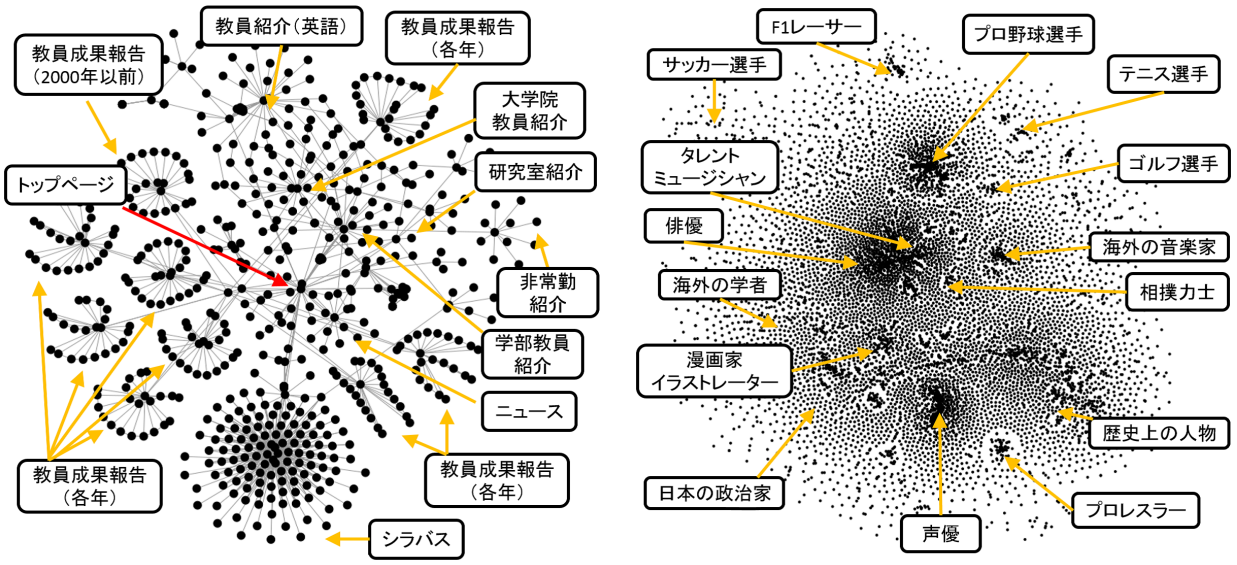
3 つ目のネットワークは，レシピ投稿サイト Cookpad におけるユーザのつくれぽ関係である [*3]．ユーザをノード，つくれぽ関係をリンクとした有向ネットワークを構築し，最大強連結成分を抽出，無向化した．さらにつくれぽ関係が 10 以上あるような関係のみを抽出した．各ノードの特徴ベクトルは，投稿したレシピに使用する食材の使用頻度とした．ノード数は 7,815，リンク数は 40,569，特徴ベクトルの次元数は 4,171 である．本稿では Cookpad ネットワークと呼ぶ．

6.2 推定パラメータに関する考察

提案手法において距離減衰重みを制御するパラメータを推定した結果 $\hat{\lambda}$ について考察する．本稿では，パラメータの推定値によってランキングし，上位，下位のノードの特徴について定性的に述べる．

Web ネットワークにおいて，$\hat{\lambda} > 10$ となるようなランキング上位は，「教員紹介ページ」が多くを占めていた．これ

[*1] 法政大学情報科学部（2010 年 8 月時点）http://cis.k.hosei.ac.jp/
[*2] https://ja.wikipedia.org/
[*3] http://cookpad.com/



(a) Web ネットワーク　　(b) Wiki ネットワーク

図 1　正解ラベル

らのページには，学生向けに教員の経歴や研究内容などが書かれており，隣接する他の教員ページも類似の単語（名詞）を使用した内容になっている．さらに，他の教員ページへは少ないステップでたどり着くことができ，類似の内容のページが近隣に集まり，逆に遠くには類似のページが存在しないため，パラメータの値が大きくなり，強い距離減衰重みを実現させたと考えられる．$\hat{\lambda} \simeq 0$ となるようなランキング下位は「ニュースページ」や「お知らせページ」が多くを占めていた．これらのページには，所属学生や教員の受賞ニュースなどが書かれており，近隣はおろかネットワーク内に類似するページが存在しないため，パラメータの値が小さくなり，弱い距離減衰重みを実現させたと考えられる．

Wiki ネットワークにおいて，$\hat{\lambda} > 10$ となるようなランキング上位は，「タレント芸能人ページ」が多くを占めていた．これらはバンドやコンビ，チームなどのメンバーになっている率が高く，共起関係にある（同じチームに所属）ノード同士は，類似の単語を使用する傾向にあるため，このような結果になったと考えられる．すなわち，タレントの共起関係は，類似の特徴ベクトルをもったノード同士が結びついているという直感に合致した結果となった．一方，「俳優や女優などの芸能人ページ」は，タレント芸能人と比べると，パラメータの値は低めに推定された．ウィキペディアの芸能人ページ内には，来歴・人物やエピソードなどが書かれているが，共起関係にある俳優同士（同じドラマに出演など）だからといって，これらに用いられる単語が類似するとは限らないため，それを反映した結果と考えられる．すなわち，俳優・女優の共起関係は，様々な特徴ベクトルをもったノード同士が結びついているという直感に合致した結果となった．$\hat{\lambda} \simeq 0$ となるようなランキング下位は，「特異な固有名詞を使用するページ」や「内容が少ないページ」などが見られた．これらのページは，ネットワーク内に類似するページが存在しないため，距離に関係なく様々なページの特徴ベクトルを合成することで，自身の特徴ベクトルとのコサイン類似度を高くしようとする．そのためパラメータの値が小さくなり，弱い距離減衰重みを実現させたと考えられる．

Cookpad ネットワークにおいて，$\hat{\lambda} > 10$ となるようなランキング上位は，「比較的小規模なカテゴリコミュニティに属するユーザ」が多くを占めていた．具体的には，「タレ」，「幼児食」，「ドリンク」などがあげられる．これらはカテゴリに投稿するユーザは，比較的小規模ながらコミュニティを形成しており，コミュニティ内では類似の食材を利用するユーザが多く存在する．一方で，コミュニティの外には類似するユーザがあまり存在しないため，距離減衰重みを制御するパラメータの値が大きく推定されたと考えられる．$\hat{\lambda} \simeq 0$ となるようなランキング下位は，「様々なカテゴリのレシピを投稿するユーザ」が多くを占めていた．様々な食材を使用するため，近隣だけでなく遠方のノードの特徴ベクトルをも合成しなければ高いコサイン類似度を得ることができない．いわば，コミュニティにあまり染まらないノードである．そのため，距離減衰重みを制御するパラメータの値が小さく推定されたと考えられる．これらの結果から，推定パラメータ $\hat{\lambda}$ の値は，コミュニティのサイズや所属するノードの特徴ベクトルのばらつき度に依存することが示唆された．

次に，推定値 $\hat{\lambda}_u$ と次数 d_u，他のノードへの距離最大値 D_u，実現できた目的関数値（コサイン類似度）の間の相関係数を表 1 に示す．このように，ネットワーク構造から定量化される特徴量とは大きな相関がなかった．裏を返

表 1　パラメータ推定値とネットワーク指標の相関

	d_u	D_u	$F(\hat{\lambda}_u)$
WebNW	3.222075e-1	-2.305069e-1	-6.286301e-2
WikiNW	1.926421e-1	-8.677814e-2	3.874408e-1
CookpadNW	1.641944e-1	-2.460673e-2	-7.678937e-2

せば，ノードの特徴ベクトルとネットワーク構造の両方を考慮することで，コミュニティに順応しているか，外れ値的な存在なのかを示す新たな中心性指標と捉えることもできる．

6.3　アノテーション付与結果

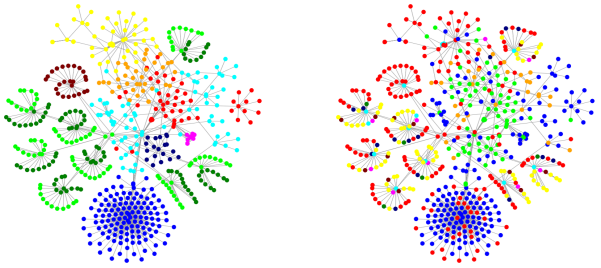
図 1 に，Web ネットワークと Wiki ネットワークの正解ラベルを示す．可視化にはノード間の隣接関係により布置座標を計算するクロスエントロピー法 [7] を用いる．図 2 に，$K = 10$ とした際の Web ネットワークに対するノード分割結果を示す．可視化結果からわかることとして，(b) では，隣接するノードでも異なるクラスタ（色）が割り当てられており，本稿の目的である隣接関係を考慮できていない．(c) では，ネットワーク構造上綺麗に分割できているが，意味的な部分（特徴ベクトルの類似性）で分割されている保証はない．(d) では，可視化座標の近接性に依存しているため，(c) のコミュニティ抽出とも異なる結果が得られた．また，(c) と (d) に対する Z スコア上位の特徴量には共通点がなく，アノテーションには不向きな特徴量であった．これらと比較して推定パラメータを用いた (a) の提案手法では，隣接関係を考慮しているため，連結したノード群単位で同一のクラスタに割り当てられており，かつ，意味的に類似するノードが同一のクラスタに割り当てられている．実際にアノテーションとして抽出された特徴量を表 2 に示す．正解ラベルと図 2(a) のノードの色を念頭に置いて見ると，どのクラスタに付されたアノテーション特徴量も，ある程度クラスタに属するノードの特色を表すものが抽出されている．特に，CNM クラスタリングでは「教員成果報告ページ」として同一クラスタに分けられていたノード群が，提案手法では第 2 クラスタのような画像処理系の教員成果報告ページと，第 9 クラスタのような Web 系の教員成果報告ページに分けられている．提案手法は，意味的なまとまりを考慮するため，近隣に存在していても特徴ベクトルが大きく異なれば分離することが可能である．

図 3 に，$K = 10$ とした際の Wiki ネットワークに対するノード分割結果を示す．可視化結果からわかることとして，(b) では，隣接するノードでも異なるクラスタ（色）が割り当てられており，本稿の目的である隣接関係を考慮できていない．(c)(d) では，ネットワーク構造上綺麗に分割できているが，意味的な部分で分割されている保証はない．しかし，共起ネットワークの性質上，ある程度の意味的なまとまりのあるノード群が抽出されたように見受けれる．これらと比較して推定パラメータを用いた (a) の提案手法では，(b) の意味的まとまりと (c) の構造的まとまりの両方を考慮できているように見える．抽出されたアノテーション特徴量を表 3 に示す．正解ラベルと図 3(a) のノードの色を念頭に置いて見ると，どのクラスタに付されたアノテーション特徴量も，ある程度クラスタに属するノードの特色を表すものが抽出されている．特に，CNM クラスタリングでは「歴史上の人物」と「政治家」が同一クラスタに分けられていたノード群が，提案手法では第 1 クラスタの「歴史上の人物」と第 6 クラスタの「政治家」に分けられている．提案手法は，周辺ノードの特徴ベクトルを合成するため，離れたところに存在する類似ノード群を分離することも可能である．

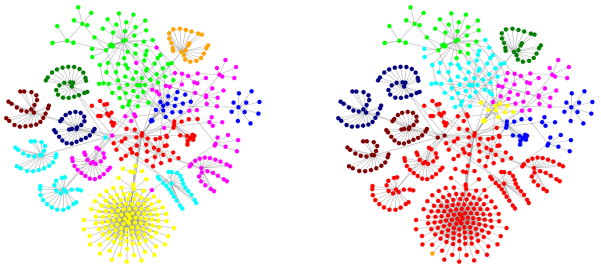
また，紙面の都合上 Cookpad ネットワークの可視化結果は掲載していないが，概ね上記 2 つのネットワークの結果と矛盾のない結果が得られた．アノテーション特徴量を表 4 に示す．こちらの結果も，「スイーツ」の食材が多い第 1 クラスタ，「朝食料理」の食材が多い第 4 クラスタ，「つけもの」の食材が多い第 7 クラスタ，「お弁当」の食材が多い第 8 クラスタ，「ケーキ」の食材が多い第 9 クラスタ，「カレー」の食材が多い第 10 クラスタなど，アノテーションとして有用な特徴量が抽出されている．

次に，比較手法と提案手法を定量的に比較する．あるクラスタ（ノード群）に有意に出現する特徴量があれば，そのクラスタに意味的なまとまりがあると言える．抽出したクラスタに意味的なまとまりがあるか，Z スコアの意味で定量的に評価した結果を図 4 に示す．実際には，各クラスタの上位 10 件の特徴量 Z スコアの平均をプロットした．図 4 を見ると，どのネットワークにおいても，K-medoids クラスタリングが最も高い値を示している．これは，特徴ベクトルを直接クラスタリングしているので当然の結果であるが，提案手法も次いで高い値を示している．提案手法では，距離減衰合成ベクトルをクラスタリングしているため，このような結果が得られることは自明ではないことを注意しておく．反対に，対象ネットワークの構造に依存するが，CNM クラスタリングでは意味的なまとまりは見られない傾向にある．CNM クラスタリング同様，ネットワークの隣接関係が影響する MST 分割でも，意味的なまとまり度は見られない傾向にある．

次に，抽出したクラスタにネットワーク構造としてのまとまりがあるかを定量的に評価した結果を図 5 に示す．実際には，割り当てたクラスタ番号を各ノードの属性値として，Assortative 係数 [1] を計算することにより評価した．図 5 を見ると，ノードの隣接関係のみを考慮している CNM クラスタリングが最も高い値を示しているが，この結果は自明である．次いで提案手法も高い値を示しており，同一クラスタのノード同士が隣接関係にあることが，可視化結果からだけでなく定量的にも示された．反対に，隣接関係

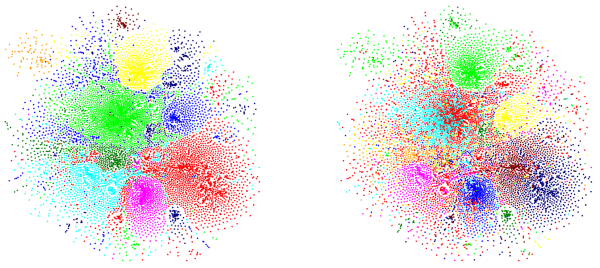


(a) 提案手法　(b) K-medoids クラスタリング

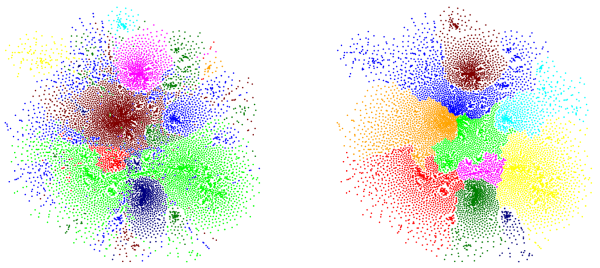


(c) CNM クラスタリング　(d) MST 分割

図 2　Web ネットワークノード　分割結果（$K = 10$）



(a) 提案手法　(b) K-medoids クラスタリング



(c) CNM クラスタリング　(d) MST 分割

図 3　Wiki ネットワーク　ノード分割結果（$K = 10$）

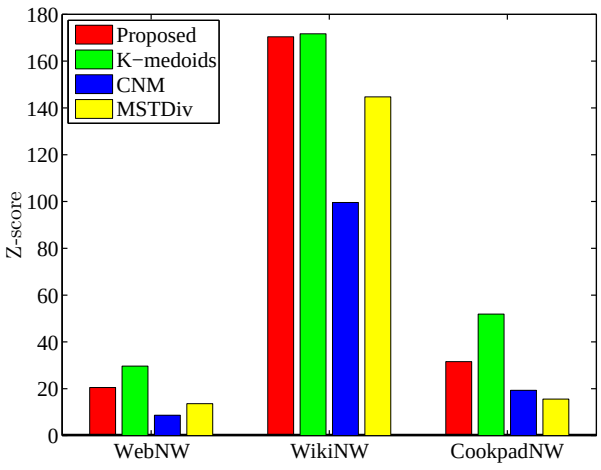


図 4　意味的まとまり度の定量評価

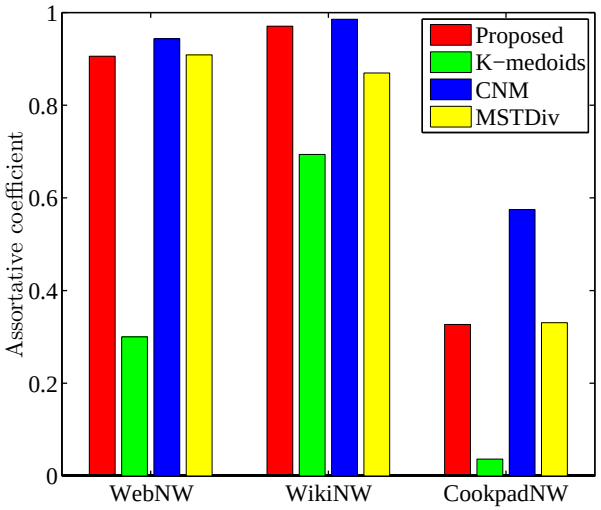


図 5　構造的まとまり度の定量評価

を一切考慮していない K-medoids クラスタリングは最も低い値を示している．これらの結果より，提案手法は本稿の目的である，隣接関係を考慮した特徴的な意味を有するノード群を抽出できていることがわかった．

7. おわりに

本稿では，距離減衰重みを導入した合成ベクトルを考え，ノード間の隣接関係を考慮した，ノード群へのアノテーション法を提案した．距離減衰を実現するに際し，各ノードごとに異なる周辺ノードとの親和性や異端児度を意味するパラメータを数理的枠組みのもとで推定した．実ネットワークを用いた評価実験より，推定パラメータは，ある程度妥当にノードの異端児度を示していることを確認した．さらに，推定パラメータを用いて距離減衰重み付き合成ベクトルを構築し，アノテーション付与を実現した．アノテーション法としての評価では，K-medoids クラスタリングで抽出できる意味的なまとまりと CNM クラスタリングで抽出できる構造的なまとまりの両側面を併せ持っており，隣接関係にあるノード群に対し適切なアノテーションを付与できることが示唆された．今後は，人工データを用いて提案手法の精度などを定量的に評価するとともに，大規模ネットワークでも耐えられるようにアルゴリズムの高速化にも着手する予定である．

謝辞　本研究は，JSPS 科研費 (No.15J00735) の助成を受けたものである．本研究の評価に際し，クックパッド株式会社と国立情報学研究所が提供する「クックパッドデータ」を利用した．ここに記して謝意を示す．

表 2　Web ネットワークのアノテーション特徴量

	ノードの色	1 位	2 位	3 位	4 位	5 位
クラスタ 1	#ff0000	科学	情報	研究	コンピュータ	学科
クラスタ 2	#00ff00	node	algorithm	image	virtual	convert
クラスタ 3	#0000ff	科目	授業	理解度	春	秋
クラスタ 4	#ffff00	research	year	student	advisor	English
クラスタ 5	#ff00ff	課程	セミナー	指導	単位	博士
クラスタ 6	#00ffff	画像	映像	描画	認識	動画
クラスタ 7	#ffa500	領域	研究	開発	プロジェクト	非常勤
クラスタ 8	#800000	page	proceeding	transaction	press	edition
クラスタ 9	#008000	model	browser	object	agent	function
クラスタ 10	#000080	掲載	受賞	時間割	更新	開催

表 3　Wiki ネットワークのアノテーション特徴量

	ノードの色	1 位	2 位	3 位	4 位	5 位
クラスタ 1	#ff0000	旧暦	幕府	徳川	藤原	元年
クラスタ 2	#00ff00	テレビ	出演	フジテレビ	ドラマ	番組
クラスタ 3	#0000ff	交響	ピアノ	音楽	フランス	ローマ
クラスタ 4	#ffff00	野球	選手	プロ	投手	本塁打
クラスタ 5	#ff00ff	アニメ	声優	ガンダム	戦士	ロボット
クラスタ 6	#00ffff	内閣	議員	選挙	大臣	大統領
クラスタ 7	#ffa500	サッカー	ワールドカップ	得点	代表	リーグ
クラスタ 8	#800000	グランプリ	ドライバー	レース	モナコ	フェラーリ
クラスタ 9	#008000	漫画	連載	文庫	作品	手塚
クラスタ 10	#000080	場所	優勝	王座	オープン	プロレス

表 4　Cookpad ネットワークのアノテーション特徴量

	1 位	2 位	3 位	4 位	5 位
クラスタ 1	薄力粉	砂糖	牛乳	強力粉	マーガリン
クラスタ 2	オリーブオイル	塩	ニンニク	白ワイン	植物油
クラスタ 3	酒	コショウ	醤油	だし汁	ごま油
クラスタ 4	食パン	E マフィン	マヨネーズ	ベーコン	卵
クラスタ 5	醤油	みりん	麺つゆ	酒	だし汁
クラスタ 6	バニラオイル	無塩バター	全粒粉	上白糖	牛乳
クラスタ 7	塩麹	きゅうり	ナス	大根	甘酢
クラスタ 8	海苔	ご飯	チーズ	ハム	ウィンナー
クラスタ 9	グラニュー糖	生クリーム	卵黄	薄力粉	無塩バター
クラスタ 10	じゃがいも	ウスターソース	ルー	玉ねぎ	カレー粉

参考文献

[1] Newman, M. E. J.: Assortative mixing in networks, *Structure*, Vol. 2, No. 4, p. 5 (online), DOI: 10.1103/PhysRevLett.89.208701 (2002).

[2] Kuramochi, T., Okada, N., Tanikawa, K., Hijikata, Y. and Nishida, S.: Community Extracting Using Intersection Graph and Content Analysis in Complex Network, *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '12, Vol. 1, Washington, DC, USA, IEEE Computer Society, pp. 222–229 (2012).

[3] Wu, Y., Jin, R., Zhu, X. and Zhang, X.: Finding Dense and Connected Subgraphs in Dual Networks, *Proceedings of the IEEE 31st International Conference on Data Engineering (ICDE2015)*, pp. 915–926 (2015).

[4] 小林えり，斉藤和巳，池田哲夫，大久保誠也：L1 埋め込みによるアノテーション付き可視化法，第 7 回 Web とデータベースに関するフォーラム (WebDB Forum2014) (2014).

[5] Clauset, A., Newman, M. E. J. and Moore, C.: Finding community structure in very large networks, *Physical Review E*, Vol. 70, No. 6, pp. 066111+ (online), DOI: 10.1103/PhysRevE.70.066111 (2004).

[6] Vinod, H.: *Integer Programming and The Theory of Grouping*, Vol. 64, An Official Journal of the American Statistical Association (1969).

[7] Yamada, T., Saito, K. and Ueda, N.: Cross-entropy directed embedding of network data, *Proceedings of the 20th International Conference on Machine Learning (ICML03)*, pp. 832–839 (2003).