構造特性と意味特性を考慮した中心性指標の提案

伏見　卓恭†　　佐藤　哲司†　　斉藤　和巳††　　風間　一洋†††

† 筑波大学図書館情報メディア系　〒 305-8550 茨城県つくば市春日 1-2
†† 静岡県立大学経営情報学部　〒 422–8526 静岡県静岡市駿河区谷田 52-1
††† 和歌山大学システム工学部　〒 640–8510 和歌山県和歌山市栄谷 930 番地
E-mail: †{fushimi,satoh}@ce.slis.tsukuba.ac.jp, ††k-saito@u-shizuoka-ken.ac.jp, †††kazama@ingrid.org

あらまし　本研究では，ネットワークにおけるノードの活動履歴や性質などから得られるコンテンツベクトルを用いて，コンテンツ中心性という新たな中心性指標を提案する．ネットワークでは，類似のコンテンツを有するノード同士は偏って分布しており，分布の中心に存在するノードからの距離が離れるほど，コンテンツ密度は徐々にあるいは急激に減少すると想像できる．そこで各ノードに対して，自身からの距離に従って減衰する重みを付しながらコンテンツベクトルを合成し，自身のコンテンツベクトルとのコサイン類似度により各ノード周りのコンテンツの集中度を定量化し，ランキングする．3 つの実ネットワークを用いた評価実験では，中心性ランキングの妥当性を確認するとともに，減衰レベルを調整するパラメータの推定結果とノードの性質，想定される分布の性質を比較しながら考察する．さらに，特定のノード群において有意に多く出現するコンテンツを用いてノード群にアノテーションを付与することで，コンテンツ分布の凝集性を評価する．

キーワード　中心性，ネットワーク，構造特性，意味特性

1. はじめに

Twitter や Facebook 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では，ユーザ間に多くのインタラクションが存在し，ネットワークとして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，フォロー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のベクトルが類似していることは，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変化点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様々である．一方，ネットワークを構成する各ノードに対して，構造特性からノードをランキングする指標として，社会ネットワーク分析で用いられる中心性指標があげられる [2]．代表的なものとして，他ノードとの隣接度合いに着目した次数中心性，他ノードとの距離に着目した近接中心性，任意のノードペア間を媒介する度合いに着目した媒介中心性，隣接ノードの中心性を加味して自身の中心性を再帰的に求める固有ベクトル中心性，所属するコミュニティへの帰属度に着目したコミュニティ中心性 [3] などが広く知られている．さらには，Web ページのランキング手法である，PageRank や HITS なども中心性としての役割を果たす．これらの中心性指標は，構造特性のみに着目しているため，ノードの活動履歴から得られる特徴を十分に反映出来ていない．すなわち，ノード A とノード B が隣接関係にあっても，これらのノードがどの程度類似しているかは示していない．

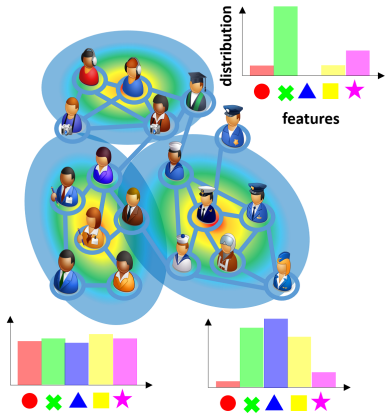


図 1　ネットワーク上でのコンテンツ分布

本研究では，ネットワーク構造上のコンテンツ分布のモード（最頻値）を発見するために，コンテンツの集中度を定量化する（図 1 参照）．分布のモードを発見する特徴空間分析の手法として，カーネル密度推定の技術を用いた MeanShift クラスタリングがあるが，ネットワーク構造上の分布に着目した手法は存在しない．ネットワーク上でのコンテンツ分布を想定して，各ノードごとにその近辺にコンテンツが偏在する度合いを表すコンテンツ中心性という新たな中心性指標を提案する．具体的には，各ノードの特徴量をコンテンツベクトルで表現し，近隣ノードのコンテンツベクトルを合成したベクトルとのコサイン類似度により類似度を定義する．ただし，直接隣接するノードだけでなく，数ノードを介して存在するノードのコンテンツベクトルも考慮する．この際，どの程度離れたノードまで合成すべきかは自明ではない．また一般に，離れたノードの影響も少

なからず受けるが，その影響は遠いほど小さくなると考えるのが自然である．そこで，ノード間の距離に応じて減衰する重みを乗じながら合成ベクトルを構築していく [4]．減衰度合いは各ノードによって異なるが，近隣に狭く類似ノードが存在するときは減衰を強くし，近隣に広く類似ノードが存在するときは減衰を弱くする．すなわち，各ノードに対して，適切な減衰を実現するパラメータを推定する必要がある．本研究では，コサイン類似度が最大になるように各ノードの減衰パラメータを推定し，そのパラメータの元でコンテンツベクトル間のコサイン類似度を計算し，各ノードのコンテンツ中心度とする．推定されたパラメータの値は，各ノードが特徴量を共有する近隣ノードの範囲を規定する意味もある．近隣ノードの定義として，CNM 法 [5] などにより抽出されたコミュニティを用いることもできるが，コミュニティの境界がコンテンツの境界である保証はない．また，コンテンツの境界が厳密に存在するわけでもないため，減衰させることにより連続的に表現する．

2. 関 連 研 究

中心性指標に関する多くの先行研究が存在するが，それらの多くはネットワーク構造のみに着目している．著者らの知る範囲では，本稿で提案するようなネットワーク構造上のコンテンツ密度による中心性指標を提案した論文は存在しない．一方，コミュニティ抽出に関する先行研究では，構造特性と意味特性の両方に基づく手法がいくつか提案されている．そこで，ノードのコンテンツを用いるコミュニティ抽出の関連研究について説明する．

Kuramochi ら [6] 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 [7]．この手法では，低次数ノードを枝刈りすることで構造的に密な部分を抽出し，効率的なアルゴリズムを実現している．本稿の提案指標では，構造的な密度ではなく意味的な密度に着目しており，全ノードの中心性スコアを計算する点で異なる．

これらの関連研究と異なり，我々の手法では，自身のコンテンツベクトルと近隣ノードのコンテンツベクトルを合成したベクトル間のコサイン類似度により各ノードの中心性スコアを計算する．

3. 提 案 手 法

この節では，本稿の提案指標である**コンテンツ中心性**について説明する．コンテンツ中心性は，類似コンテンツを持つノードがそのノード近隣にどの程度偏在しているか，すなわち，各ノード周辺のコンテンツ集中度を各ノードの中心性スコアとして定量化し，ノードをランキングする．まず，コンテンツ中心性の概念および仮定について説明し，中心性スコアの計算法および，減衰パラメータの推定法について述べる．

3.1 概念と仮定

現実のネットワークにおいて，ソーシャルメディアでの投稿内容などノードの活動履歴から得られる特徴量は，隣接関係にあるノード同士で類似する傾向が観測されている [1]．本研究では，こうした類似の特徴量を有するノード群は，ネットワーク上に偏在していると仮定する．すなわち，コンテンツに関して隣接ノード間の類似性が高い（同類選択性が高い）ネットワークを対象とする．さらに，特徴量分布の密度は分布の中心に位置するノードとの距離が離れるにつれて，徐々にあるいは急激に減衰すると仮定する（図 2）．そこで各ノードに対して，自身とその近隣に類似コンテンツがどの程度集中しているかを定量化することを試みる．具体的には，投稿内容の TF・IDF などでコンテンツベクトルを定義し，距離に従って減衰する重みを付しながら，近隣ノードのコンテンツベクトルを合成する．自身のコンテンツベクトルと近隣ノードの合成ベクトル間のコサイン類似度により集中度を定量化する．前述のように，隣接関係にあるノード同士は類似のコンテンツを有する，すなわち，互いに強い影響を与えていると言える．逆に，遠く離れたノードのコンテンツはほとんど影響しないと自然に想像できる．この影響を反映するために，図 2 のようにノード間の距離に基づく減衰関数を導入する．

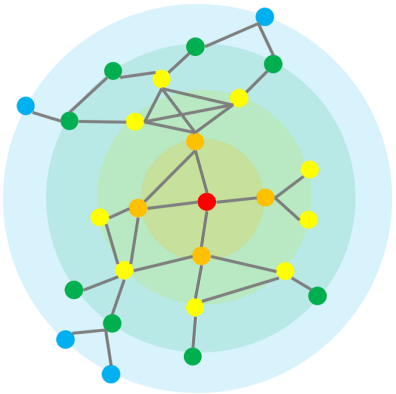


図 2　コンテンツ分布と影響度（中心：赤ノード）

本研究では，コンテンツの集中度をコンテンツ中心性として定義する．中心性の高いノードの近隣には，そのノードのコンテンツと類似のコンテンツを有するノードが偏在していることを意味する．

また，ネットワーク全体を俯瞰すると，類似のコンテンツをもつノード群が，1 箇所だけではなく全く離れた場所にも存在する場合も想定できる．このように距離的に離れて存在してい

る場合は，別の分布（多峰性）として扱うことができる．

3.2 中心性スコア計算

ノード集合 V とリンク集合 E からなる単純無向ネットワーク $G=(V,E)$ の各ノード $u \in V$ は，J 次元コンテンツベクトル $\mathbf{x}_u$ を有する．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．ノード u の距離 d である近隣ノード集合を $\Gamma_d(u)=\{v : d(u,v)=d\} \subset V$ とする．コンテンツ分布の仮定より，離れたノードはほとんど影響しないようにするために，2 つの減衰関数を導入する．1 つ目は，以下に定義する指数的減衰関数であり，

$$\rho(d;\lambda)=\exp(-\lambda d),$$

λ は減衰の程度を制御するパラメータである．2 つ目は，以下に定義するべき乗的減衰関数である：

$$\rho(d;\lambda)=\exp(-\lambda \log d).$$

各ノードに対して，距離に基づく減衰重みを付しながら，隣接するノードのコンテンツベクトルを以下のように合成する：

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

ここで，$D_u=\max_{v\in V} d(u,v)$ であり，この合成ベクトルを RVwD（Resultant Vector with Decay）と表記する．各ノードの RVwD は，直接隣接するノードを含め，近隣ノードのコンテンツベクトルを大きい重みで，遠方ノードのコンテンツベクトルを小さい重みで合成したものである．したがって，RVwD は幾分か均されている．

次に，各ノードに対して，元々のコンテンツベクトルと RVwD 間のコサイン類似度によりコンテンツ中心性のスコアを計算する：

$$\mathrm{CDC}(u)=\langle \mathbf{x}_u, \mathbf{y}_u\rangle = \left\langle \mathbf{x}_u, \frac{\sum_{d=1}^{D_u}\rho(d;\lambda)\sum_{v\in\Gamma_d(u)}\mathbf{x}_v}{||\sum_{d=1}^{D_u}\rho(d;\lambda)\sum_{v\in\Gamma_d(u)}\mathbf{x}_v||}\right\rangle. \quad (2)$$

ここで，$\mathbf{x}_u$ は L2 ノルムが 1 になるように正規化してある．ノード u のスコア $\mathrm{CDC}(u)$ が他のノードより大きければ，u の界隈に類似コンテンツを有するノードが偏在しており，ノード u はコンテンツ中心性ランキングで上位となる．

3.3 減衰パラメータ推定

上述した RVwD $\mathbf{y}_u$ は，近隣ノードのコンテンツベクトルを減衰させながら合成させて構築する．ネットワーク的に近いノードほどコンテンツベクトルが類似する傾向にあるという前提に従って，ノードごとに減衰度合いを調整する．本稿では，各ノードの RVwD がコンテンツベクトルとの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．L2 ノルムを 1 に正規化したコンテンツベクトルを $\mathbf{x}_u$ とし，以下の目的関数を定義する：

$$F_u(\lambda_u)=\mathbf{x}_u^T \frac{\sum_{v\in V\setminus\{u\}}\rho(d(u,v);\lambda_u)\mathbf{x}_v}{||\sum_{v\in V\setminus\{u\}}\rho(d(u,v);\lambda_u)\mathbf{x}_v||}. \quad (3)$$

目的関数 (3) 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．ここで，d を $\log(d)$ と置き換えることで，以下で説明する導出はべき乗的減衰でも成り立つため，以下では指数減衰重みを用いて説明を進める．ノード u に対して，距離 d にあるノードのコンテンツ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 のコンテンツ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 (3) 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：

$$\begin{aligned}\log F_u(\lambda_u) &= \log\sum_{d=1}^{D_u}\exp(-\lambda_u d)g_{u,d} \\ &\quad -\frac{1}{2}\log\sum_{d=2}^{2D_u}\exp(-\lambda_u d)h_{u,d}. \end{aligned} \quad (4)$$

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 (4) は以下のように書き換えられる：

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

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強い減衰を実現し，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 に近いほど弱い減衰を実現し，近隣も遠方も同程度の重みで合成する．すなわち，近隣に類似のコンテンツ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，多くのノードのコンテンツ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．

コンテンツベクトルの次元を J，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 $O(|V|\times 2\bar{D}\times J)$ である．本提案指標では，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．

4. 評 価 実 験

4.1 ネットワークデータ

本研究では，コンテンツに関して隣接ノード間の類似性が高い（同類性が高い）ネットワークを対象として評価する．1 つ目のネットワークは，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 (注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．各ノードのコンテンツベクトルは，ウェブページの内容を形態素解析して得られる名詞群の Bag of Words とした．ノード数は 600，リンク数は 1,299，コンテンツ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 (注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．各ノードのコンテンツベクトルは，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，コンテンツ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 (注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．各ノードのコンテンツベクトルは，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，コンテンツベクトルの次元数は 4,171 である．本稿では Recipe ネットワークと呼ぶ．

4.2 推定パラメータに関する考察

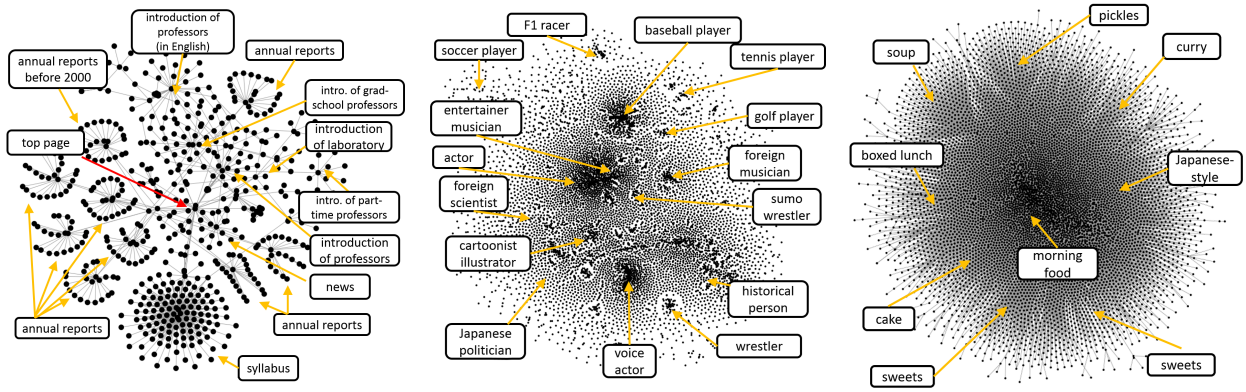
提案手法において距離減衰重みを制御するパラメータを推定した結果 $\hat{\lambda}$ について考察する．指数的減衰とべき乗的減衰によるパラメータの推定結果には高い相関があったため，指数減衰パラメータの推定値によってランキングし，上位，下位のノードの特徴を定性的に述べる．

Web ネットワークにおいて，$\hat{\lambda} > 1$ 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．これらのページには，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，$\hat{\lambda} > 1$ 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割合が高く，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，類似のコンテンツ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，様々なコンテンツ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コンテンツベクトルを合成することで，自身のコンテンツ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．

Recipe ネットワークにおいて，$\hat{\lambda} > 1$ 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

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



(a) Web ネットワーク　(b) Wiki ネットワーク　(c) Recipe ネットワーク

図 3　手動によるラベル付き可視化結果

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コンテンツ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コンテンツベクトルのばらつき度に依存することが示唆された．

4.3　中心性ランキングに関する考察

この節では，提案指標であるコンテンツ中心性ランキングの結果について，上位ノードの性質や他の中心性指標と比較しながら考察する．表 1 に，各ネットワークでの上位 10 ノードを示す．Web ネットワークでは，上位 10 ノードの全てがシラバスページであり，実際にこれらのページ群には類似の単語（名詞）が多く含まれており，シラバスページ群が大きなコンテンツ分布を構成していることが伺える．Wiki ネットワークでは，ほとんどすべての上位ノードがジャニーズ事務所所属のアイドルに関するページであり，同一グループに所属するアイドルのページは互いにつながっており，かつ，ページ内に含まれる名詞も類似のものが多い．コンテンツ中心性は，類似のコンテンツベクトルを有するノードが多く分布する部分を抽出している．Recipe ネットワークに対しては，プライバシーの都合上ノードの名前を非表示にしている．このネットワークでは，ほとんどすべての上位ノードが “スープ”，“お弁当”，“カレー” に関するレシピを多く投稿するユーザコミュニティに属している．これらの各コミュニティでは，各料理で類似の具材を使用する傾向にある（例：ジャガイモ，ニンジン，カレー粉，タマネギがほとんどすべてのカレーで使用される．）．そのため，それらの具材（コンテンツ）が共起する部分を抽出できていると言える．

コミュニティ抽出法を用いれば，密につながるノード群を抽出できる．しかしながら，類似のコンテンツベクトルを有するノード群を抽出できる保証はない．一方，提案指標のように各ノードに対してコンテンツ分布の集中度を計算することで，コンテンツ分布の中心に存在するようなノードを検出できる．

表 2　中心性間の順位相関係数

	WebNW	WikiNW	RecipeNW
次数中心性	0.47	0.52	0.19
近接中心性	0.66	0.73	0.43
媒介中心性	0.28	0.19	0.30
PageRank	0.49	0.53	0.22
固有ベクトル中心性（HITS）	0.27	0.15	0.20
コミュニティ中心性	0.76	0.78	0.53
コンテンツベクトルの次元	0.14	0.02	0.18

次に，従来の構造に基づく中心性指標との関係を評価する．表 2 に，スピアマンの順位相関係数の結果を示す．評価実験に用いた 3 つ全てのネットワークにおいて，以下の関係が見られた．

- 近接中心性とコミュニティ中心性はコンテンツ中心性とある程度の相関がある．これは，コンテンツ分布のモードはコミュニティなどの中心に位置していることから，近接中心性，コミュニティ中心性が高いノードであるという直感に合致した結果である．
- 次数中心性，PageRank はやや相関がある．次数中心性と PageRank の間には強い相関関係があるが，次数が高いノードは，相対的に多くのコンテンツベクトルを強い重みで合成するため，コサイン類似度も相対的に高くなるためと考えられる．
- 媒介中心性，HITS，コンテンツベクトルの次元は，相対的に相関係数が低い傾向にある．

5.　コンテンツの凝集性に関する評価

RVwD は近隣ノードの特徴量を含んでいるベクトルであるため，隣接ノードの RVwD 間の類似度は相対的に高くなるはずである．RVwD でノードをクラスタリングし得られるノードグループは，ノード同士が連結しており，かつ意味的に類似するノード群になっていると期待できる．このような性質を有するノード群を抽出できることを確認するため，複数のクラスタリング結果と比較する．これは，コンテンツの凝集性を確認する意味も含まれている．

表 1　コンテンツ中心性ランキング

	Web ネットワーク			Wiki ネットワーク			Recipe ネットワーク		
順位	スコア	ノード	ラベル	スコア	ノード	ラベル	スコア	ノード	ラベル
1	0.9927	科学技術計算 2	シラバス	0.9589	風間俊介	アイドル	0.9470	hidden	スープ
2	0.9924	型システムと関数型言語	シラバス	0.9400	東新良和	アイドル	0.9427	hidden	スープ
3	0.9923	データベース入門	シラバス	0.9378	倖田來未	ミュージシャン	0.9329	hidden	お弁当
4	0.9920	人工知能の応用	シラバス	0.9342	生田斗真	俳優	0.9325	hidden	お弁当
5	0.9919	人工知能入門	シラバス	0.9310	松本潤	アイドル	0.9281	hidden	スープ
6	0.9908	自然科学の基礎	シラバス	0.9297	大野智	アイドル	0.9280	hidden	カレー
7	0.9905	プログラミング演習 2	シラバス	0.9265	手越祐也	アイドル	0.9273	hidden	スープ
8	0.9905	テクニカルライティング 2	シラバス	0.9261	山下智久	アイドル	0.9265	hidden	スープ
9	0.9905	科学英語 2	シラバス	0.9204	亀梨和也	アイドル	0.9264	hidden	スープ
10	0.9904	離散構造 2	シラバス	0.9133	井上康生	スポーツ	0.9257	hidden	カレー

5.1　比較手法：クラスタリング法

比較に際して使用する 3 つのクラスタリング手法について説明する．

5.1.1　K-medoids クラスタリング

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 $s(v,w)$ が与えられたとき，目的関数 $\mathcal{J}(P) = \sum_{v\in V} \max_{w\in P}\{s(v,w)\}$ を最大にするような代表オブジェクト集合 P を求める．$K = |P|$ 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．

類似度 $s(u,v)$ として，RVwD 間のコサイン類似度 $s(u,v) = \langle \mathbf{y}_u, \mathbf{y}_v\rangle / ||\mathbf{y}_u|| ||\mathbf{y}_v||$ を用いる場合と，元々のコンテンツベクトル間のコサイン類似度 $s(u,v) = \langle \mathbf{x}_u, \mathbf{x}_v\rangle / ||\mathbf{x}_u|| ||\mathbf{x}_v||$ を用いる場合を比較する．後者は，各ノードのコンテンツベクトル $\mathbf{x}$ をそのまま用いているため，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．

5.1.2　CNM クラスタリング

Clauset らによって提案された CNM 法 [5] は，以下に示すリンク構造に基づくモジュラリティ $\mathcal{Q} = \sum_{k=1}^{K} \left(e_{kk} - a_k^2\right)$ 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．実際には，$\Delta\mathcal{Q}$ を最大化することで高速化を図っている．この手法では，構造的なまとまりのみを対象としている．

5.1.3　MST クラスタリング

小林らによって提案された MST 分割法 [8] 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コンテンツに基づく尤度関数 $\mathcal{L} = \sum_{k=1}^{K-1} \sum_{j=1}^{J} q_j^{(k)} \log q_j^{(k)} / q^{(k)}$ 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する．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，特徴的なコンテンツ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により可視化する．

5.2　凝集性の有意性指標

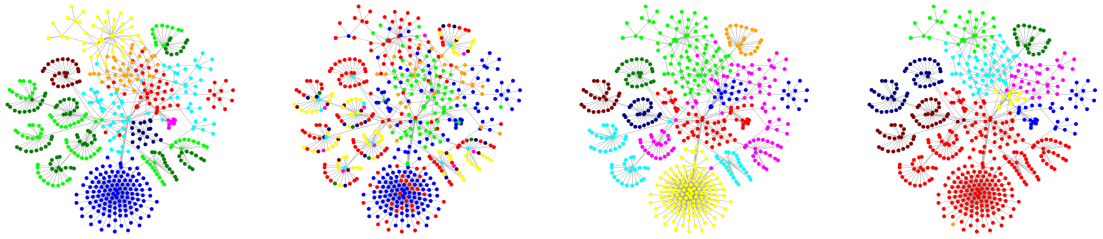
クラスタリングされたノード群（クラスタ）V_k に有意に多く出現するコンテンツを Z スコアを用いて抽出する．ここで，ネットワーク全体のコンテンツ分布を $p_j = \frac{\sum_{u\in V} x_{u,j}}{M}$ とする．分母の M は確率にするための正規化項であり，$M = \sum_{v\in V}\sum_{j=1}^{J} x_{v,j}$ である．また，クラスタ V_k に属するノードのコンテンツを $q_j^{(k)} = \sum_{u\in V_k} x_{u,j}$ のように合算する．この時，クラスタ V_k に対するコンテンツ j の Z スコアは以下のように計算する：

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

ここで，$M_k = \sum_{v\in V_k}\sum_{j=1}^{J} x_{v,j}$ である．クラスタ V_k にコンテンツ 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，各クラスタの特徴的なコンテンツを抽出する．提案手法では，各クラスタごとに Z スコア上位のコンテンツをアノテーション特徴量として採用する．

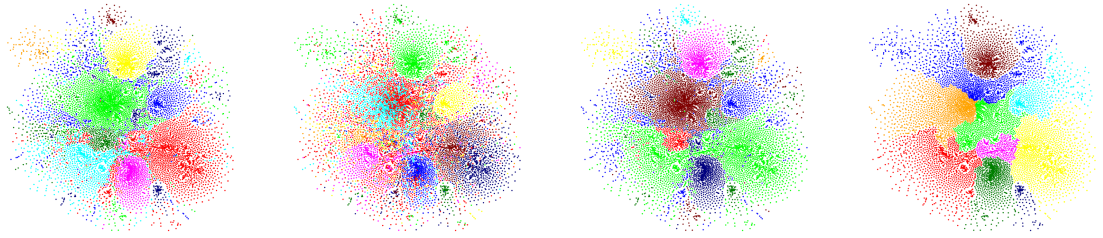
5.3　アノテーション結果に関する考察

図 3 に，各ネットワークのラベルを示す．このラベルは，著者らが手動でつけたものである．図 4 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．可視化結果から分かることとして，(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（コンテンツベクトルの類似性）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 スコア上位のコンテンツには共通点がなく，アノテーションには不向きなコンテンツであった．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特徴量として抽出されたコンテンツを表 3 に示す．Z スコアが高い順に左から表示している．図 3 のラベルと図 4(a) のノードの色を念頭に置いて見ると，どの



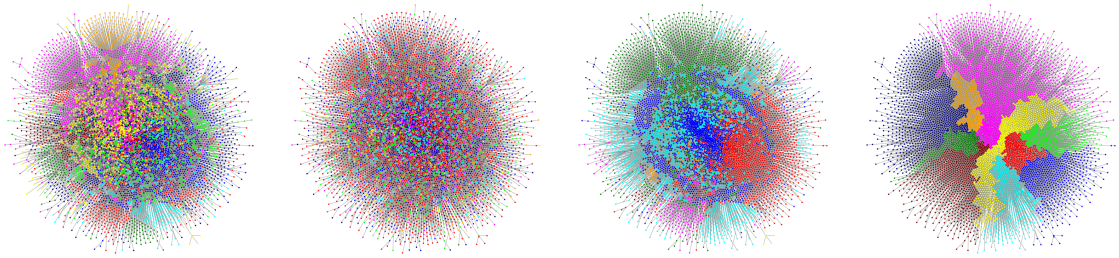
(a) Clustering by RVwDs　(b) K-medoids clustering　(c) CNM clustering　(d) MST cutting

図 4　Web ネットワークのクラスタリング結果：$K = 10$



(a) Clustering by RVwDs　(b) K-medoids clustering　(c) CNM clustering　(d) MST cutting

図 5　Wiki ネットワークのクラスタリング結果：$K = 10$



(a) Clustering by RVwDs　(b) K-medoids clustering　(c) CNM clustering　(d) MST cutting

図 6　Recipe ネットワークのクラスタリング結果：$K = 10$

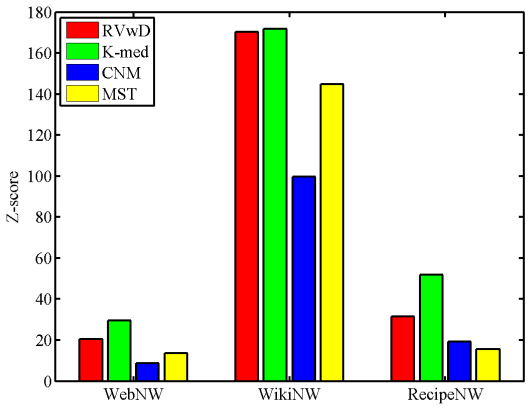


図 7　意味的まとまり度の定量評価

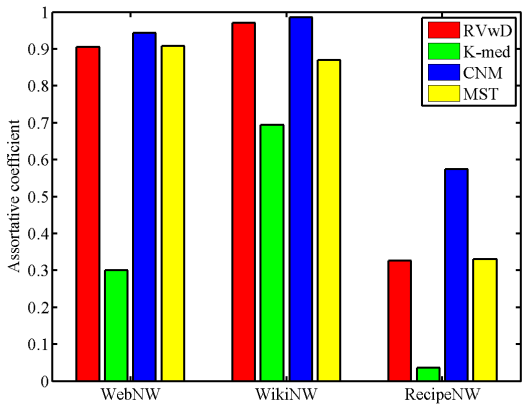


図 8　構造的まとまり度の定量評価

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コンテンツ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 5 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

表 3　アノテーション特徴量

クラスタ・色	WebNW	WikiNW	RecipeNW
1・赤	科学，情報，研究，コンピュータ，学科	旧暦，幕府，徳川，藤原，元年	薄力粉，砂糖，牛乳，強力粉，マーガリン
2・黄緑	node，algorithm，image，virtual，convert	テレビ，出演，フジテレビ，ドラマ，番組	オリーブオイル，塩，ニンニク，白ワイン，植物油
3・青	科目，授業，理解度，春，秋	交響，ピアノ，音楽，フランス，ローマ	酒，コショウ，醤油，だし汁，ごま油
4・黄	research，year，student，advisor，English	野球，選手，プロ，投手，本塁打	食パン，E マフィン，マヨネーズ，ベーコン，卵
5・桃	課程，セミナー，指導，単位，博士	アニメ，声優，ガンダム，戦士，ロボット	醤油，みりん，麺つゆ，酒，だし汁
6・水	画像，映像，描画，認識，動画	内閣，議員，選挙，大臣，大統領	バニラオイル，無塩バター，全粒粉，上白糖，牛乳
7・橙	領域，研究，開発，プロジェクト，非常勤	サッカー，ワールドカップ，得点，代表，リーグ	塩麹，きゅうり，ナス，大根，甘酢
8・茶	page，proceeding，transaction，press，edition	グランプリ，ドライバー，レース，モナコ，フェラーリ	海苔，ご飯，チーズ，ハム，ウィンナー
9・緑	model，browser，object，agent，function	漫画，連載，文庫，作品，手塚	グラニュー糖，生クリーム，卵黄，薄力粉，無塩バター
10・紺	掲載，受賞，時間割，更新，開催	場所，優勝，王座，オープン，プロレス	じゃがいも，ウスターソース，ルー，玉ねぎ，カレー粉

造的まとまりの両方を考慮できているように見える．正解ラベルと図 5(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（表 3）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，周辺ノードのコンテンツ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．

図 6 に，$K = 10$ とした際の Recipe ネットワークに対するノード分割結果を示すが，上述の 2 つのネットワークと類似の傾向が得られたため，詳細は割愛する．アノテーション特徴量についても,「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コンテンツが抽出されている（表 3）.

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）に有意に出現するコンテンツがあれば，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 7 に示す．実際には，各クラスタにおける上位 10 件のコンテンツに対する Z スコアを平均し，プロットした．図 7 を見ると，どのネットワークにおいても，コンテンツベクトルの K-medoids クラスタリング結果が最も高い値を示している．これは，コンテンツ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が，RVwD 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．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 8 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 8 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．これらの結果から，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と言える．

6. おわりに

本研究では，ネットワーク上でのコンテンツ分布を仮定して，分布の中心に存在するようなノードを抽出するコンテンツ中心性という新たな指標を提案した．3 つの実ネットワークを用いた評価実験により，ある程度妥当なノードを抽出できることを確認した．今後は，さらに多様なネットワークを用いて評価する．また，減衰の掛け方についても探求する．

謝辞　本研究は，JSPS 科研費 25280110 および JSPS 特別研究員奨励費 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．

文　献

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

[2] Freeman, L.: Centrality in social networks: Conceptual clarification, *Social Networks*, Vol. 1, No. 3, pp. 215–239 (1979).

[3] Newman, M. E. J.: Finding community structure in networks using the eigenvectors of matrices, *Physical Review E*, Vol. 74, No. 3, pp. 036104+ (2006).

[4] 伏見卓恭，佐藤哲司，斉藤和巳，風間一洋 ：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，第 8 回 Web とデータベースに関するフォーラム (WebDB Forum2015) (2015).

[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+ (2004).

[6] 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 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology*, WI-IAT '12, Vol. 1, Washington, DC, USA, IEEE Computer Society, pp. 222–229 (2012).

[7] 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).

[8] 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).