情報処理学会論文誌 Vol.58 No.6 1258–1267 (June 2017)

メトリック空間における複数カテゴリに属するハイブリッドオブジェクト抽出法の提案

伏見 卓恭[1,a)] 斉藤 和巳[2] 風間 一洋[3]

受付日 2016年10月2日, 採録日 2017年3月3日

概要：社会ネットワーク分析の分野で重要ノードを抽出する指標として媒介中心性，近接中心性が提案されており，これらは各ノードの媒介度と近接度をランキングし，上位ノードを抽出する手法である．本稿では，媒介中心性と近接中心性をベースに，メトリック空間オブジェクトの中から混合性と凝集性の高いオブジェクトを抽出する 2 つの指標を提案する．1 つ目の混合中心性は，任意のオブジェクトペアの中間に存在する度合いにより各オブジェクトの混合度を定量化することにより，データ分布のクラスタの狭間に位置するハイブリッドなオブジェクトを抽出する．2 つ目の凝集中心性は，他のオブジェクトへの距離の逆数により各オブジェクトの凝集度を定量化することにより，データ分布の凝集性が高い部分に位置するオブジェクトを抽出する．複数の実データおよび人工データを対象とした評価実験より，凝集性，混合性の高いオブジェクトを抽出できることを示す．

キーワード：中心性，メトリックデータ，Lune，代表オブジェクト

Extraction Method of Hybrid Objects Belonging to Several Categories in Metric Space

Takayasu Fushimi[1,a)] Kazumi Saito[2] Kazuhiro Kazama[3]

Received: October 2, 2016, Accepted: March 3, 2017

Abstract: In this paper, we propose two centrality measures, mixedness and cohesiveness centrality, intended to extract representative objects in a metric space. These measures are based on betweenness and closeness centrality each of which is widely used in the social network analysis field in order to extract important nodes. The mixedness centrality is an indicator that quantifies the degree to be located medway between any pairs of objects and extracts mixed objects located in medway of clusters. By contrast, the cohesiveness centrality is an indicator that quantifies the degree to be located in densely distribution and extracts cohesive objects located in center of densely distributed objects. In our experiments using several types of real and synthetic datasets, we show that our proposed centrality measures can extract characteristic representative objects in each dataset.

Keywords: centrality, metric data, Lune, representative object

1. はじめに

近年，Web 上には膨大な量のデータが蓄積されており，それらを有効活用するためにデータ間の関係や特性，代表的なオブジェクトを把握することは主要な研究課題の 1 つである．しかし一般には，これらのデータは高次元空間に分布しているため，その実態を把握することは困難である．この問題を緩和するために，多くの次元削減手法が提案されている [10], [11]．さらに，データに含まれるオブジェク

[1] 東京工科大学コンピュータサイエンス学部
School of Computer Science, Tokyo University of Technology, Hachioji, Tokyo 192–0982, Japan
[2] 静岡県立大学経営情報学部
School of Management and Information, University of Shizuoka, Shizuoka 422–8526, Japan
[3] 和歌山大学システム工学部
Faculty of Systems Engineering, Wakayama University, Wakayama 640–8510, Japan
[a)] takayasu.fushimi@gmail.com

ト群が多様体上に分布する場合も少なくない．このようなデータに対しては非線形な次元削減手法が多く提案されている [4]．

映像，音声，画像，文書，DNA 配列などのマルチメディアデータをはじめ多くのデータにおいて，オブジェクト間の距離あるいは類似度が定義できる．しかし，文字列や木構造，確率分布など，距離や類似度だけが定義されているが，ベクトルで表現できないオブジェクト群の中からある種の重要オブジェクトを抽出することは困難な場合がある．

本研究では，ベクトルで表現できるユークリッド空間のオブジェクトを一般化したメトリック空間オブジェクトを対象とする．すなわち，オブジェクト間に距離の公理を満たす距離が定義されたデータに対して，データ内に分布する重要（代表）オブジェクトを抽出する手法を提案する．ただし，重要という指標は様々な視点で変わりうる．自身と非常に類似するオブジェクトが多数存在するようなオブジェクトは，凝集性という観点で重要である．たとえば，大きな出来事に関するニュース記事は，毎日少しずつ内容を変え公開される．複数日にまたがる場合には記事数は多くなるが，内容の本質は変わらないので，それらの記事間の類似度はとても高くなる．別の視点として，大きなクラスタ間の狭間に存在するようなオブジェクトは，混合性という観点で重要である．たとえば，科学技術論文では，各専門分野の文献は大きなクラスタを形成する傾向にあるが，複数分野にまたがる学際的研究は，従来の分野の限界を飛び超え，単独の分野では解決できない問題を解決する可能性が高いと注目されている．研究者のような専門的知識を有する人材採用において，革新的な成果を期待する場合には，複数の専門分野で研究成果をあげている人物が望ましい．これらの例に限らず，凝集性と混合性のあるオブジェクトを抽出することは，現実の様々な場面に応用できる．

ソーシャルネットワーク研究の分野では，多大なノード群の中から重要なノードを抽出するための中心性という指標がいくつか提案されている [2], [3]．中心性指標として次数中心性，近接中心性，媒介中心性などが広く知られている．なかでも，他のノードへのグラフ距離の調和平均で定義される近接中心性，他のノード間の媒介回数で定義される媒介中心性は，道路ネットワークなどへの適用なども報告されており，現実問題への応用も考えられている [15], [16]．

本稿では，凝集性，あるいは，混合性が高いメトリック空間オブジェクトを抽出するために，ネットワーク中心性指標を拡張した 2 つの指標を提案する．メトリック，すなわち，オブジェクト間の距離が与えられているため，他のオブジェクトとの距離を用いて各オブジェクトの性質を定量化し，ランキングする．ネットワーク分析における近接中心性は，他のノードとの距離の総和が小さいノードを抽出する．これを拡張し，他のオブジェクトとの距離の逆数によりオブジェクトの性質を定量化することで，データ分布の凝集性が高い部分にいるようなオブジェクトを抽出するのが凝集中心性である．媒介中心性は，任意のノードペア間の最短パス上に出現する頻度を用いて各ノードの性質を定量化し，ランキングする．これを拡張し，他のオブジェクトペア間の間に存在する回数により定量化することで，データ分布のクラスタの狭間に存在するようなハイブリッドなオブジェクトを抽出するのが混合中心性である．

2 つの提案指標を評価するために，4 つの実データを用いて，メトリックデータのクラスタリング手法として著名な K-medoids 法と比較する．ランキング上位・下位のオブジェクトの特性，および，ランキング結果の違いについて評価する．

本稿の構成は以下のとおりである．2 章でネットワークを対象とした既存の媒介中心性と近接中心性について説明し，3 章でメトリック空間オブジェクトを対象とした混合中心性と凝集中心性について述べる．4 章で提案指標により抽出されたオブジェクトについて評価，考察する．5 章で既存のクラスタリング手法である K-medoids 法による代表オブジェクトとの比較，6 章で人工データを用いた頑健性の評価結果について考察し，7 章で本稿のまとめと今後の展望を述べる．

2. 既存指標

以下にネットワークを対象とした媒介中心性，近接中心性について，よく知られた文献 [13] の説明に準じて説明する．ノード集合 V，リンク集合 $E \subset V \times V$ からなる無向ネットワークを $G = (V, E)$ と表記する．文献 [13] と同様に，連結な単純無向ネットワークを前提とする．

2.1 媒介中心性

媒介中心性とは，多くのノード間の橋渡しをしているノードは重要であるという直感に基づいた指標であり，任意のノードペア間の最短パスのうち，あるノードが媒介するパスの割合によりノードをランキングする．始点ノード s と終点ノード t 間の最短パス数を $\sigma_{s,t}$，ノード v を通るノード s，t 間の最短パス数を $\sigma_{s,t}(v)$ と表す．ノード v の媒介中心性 $\mathrm{BWC}(v)$ は以下のように定義される．

$$\mathrm{BWC}(v) = \frac{\sum_{s \in V \setminus \{v\}} \sum_{t \in V \setminus \{s,v\}} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}}}{(|V|-1)(|V|-2)} \tag{1}$$

2.2 近接中心性

近接中心性とは，他の多くのノードへ少ないステップでたどり着ける，ネットワークの中心にいるノードは重要であるという直観に基づいた指標であり，任意のノードから他のノードへの距離の調和平均の逆数によりノードをランキングする．ノード v，u 間の最短パス長を $g(v, u)$ と表す．ノード v の近接中心性 $\mathrm{CLC}(v)$ は以下のように定義さ

れる．

$$\mathrm{CLC}(v) = \frac{\sum_{u \in V \setminus \{v\}} g(v,u)^{-1}}{(|V|-1)} \tag{2}$$

3. 提案指標

ネットワークにおけるノードを対象とする既存の中心性指標の概念を拡張し，メトリック空間オブジェクトに対する混合中心性，凝集中心性の概念および指標を提案する．メトリック空間のオブジェクト集合を U，任意のオブジェクト $u, v \in U$ 間のメトリックを $d(u,v)$ と表記する．

3.1 混合中心性（Mixedness centrality）

ネットワークにおける媒介度とは，他のノードペアの最短パス上に存在する回数により定義される．これを拡張して，メトリック空間において他のオブジェクトペアの間に存在する回数により各オブジェクトの混合度を定義する．オブジェクトペア間に存在するか否かについては，相対近傍グラフ [9] 構築における Lune の概念を用いる．具体的には，あるオブジェクト u が，オブジェクトペア s，t 間に存在するか否かは，s と t それぞれを中心とした半径 $d(s,t)$ の円の交わり部分（Lune）$L(s,t)$ に含まれるか否かにより判定する．オブジェクト u が Lune $L(s,t)$ 内に存在するか否かを，

$$\delta_{s,t}(u) = \begin{cases} 1 & \text{if} \quad u \in L(s,t) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

とし，オブジェクト u の混合度を以下のように定義する．

$$\mathrm{MXC}(u) = \frac{\sum_{s \in U \setminus \{u\}} \sum_{t \in U \setminus \{s,u\}} \delta_{s,t}(u)}{(|U|-1)(|U|-2)} \tag{4}$$

メトリック空間において，あるオブジェクト o が Lune $L(s,t)$ に存在するかどうかを判定する方法は，次のとおりである．まず，オブジェクト s から距離の昇順にソートし，オブジェクト o を順に取り出す．$d(s,o) > d(s,t)$ の場合は，すでに Lune の外なので終了する．次にオブジェクト o からの距離に着目し，$d(o,t) \leq d(s,t)$ の場合，オブジェクト o はオブジェクト s と t の作る Lune $L(s,t)$ に存在すると判断できる．そこで，すべてのオブジェクトを $\mathrm{MXC}(u)$ の値で降順ソートし，大きなクラスタの狭間に位置するハイブリッドな（複数カテゴリの性質をあわせ持つ）オブジェクトを抽出する．

3.2 凝集中心性（Cohesiveness centrality）

ネットワークにおける近接度とは，他のノードとの最短距離の調和平均の大きさにより定義される．これを拡張して，メトリック空間において他のオブジェクトとの距離の調和平均により各オブジェクトの凝集度を定義する．

オブジェクト u と v 間の距離 $d(u,v)$ に対し，オブジェクト u の凝集度を以下のように定義する．

$$\mathrm{CHC}(u) = \frac{\sum_{v \in U \setminus \{u\}} d(u,v)^{-1}}{(|U|-1)} \tag{5}$$

そこで，すべてのオブジェクトを $\mathrm{CHC}(u)$ の値で降順ソートし，密に分布するオブジェクト群の中に位置する（類似オブジェクトが多数存在する）オブジェクトを抽出する．

4. 評価実験

4.1 データセット

提案指標の性質と有効性を評価するために，4 つのメトリック空間オブジェクトのデータを採用する．1 つ目は，手書き文字認識用データベースに含まれるデータを 5,000 個抽出したものである．10 クラス（0，1，2，3，4，5，6，7，8，9）の数字の手書き文字データであり，各文字は $28 \times 28 = 784$ 画素で，各画素値は 0〜255 の 256 階調グレースケールで表されている．各手書き文字を，画素値を要素とする 784 次元ベクトルで表現し，文字間の類似度として，頻繁に用いられる相関係数 [8] を採用する．文字 i，j 間の相関係数 $r(i,j)$ を以下のように距離に変換し，オブジェクト間のメトリックとした：

$$d(i,j) = \sqrt{2(1-r(i,j))}.$$

距離の公理（三角不等式）を満たすように，一般的に用いられる平方根をとった距離を採用する[*1]．本稿では，文字データと呼ぶ．

2 つ目は，「東北大・松下単語音声データベース Vol.5」に含まれる 3,263 個の日本の駅名を，音素を表す記号で表記した単語データである [14]．単語間の距離として，代表的な編集距離であるレーベンシュタイン距離 [5] を用いる．これは，2 つの文字列が文字（character）の挿入・削除・置換で同一になる場合の最小の操作回数をコストとする．ただし，そのままでは長い単語は必然的に距離が大きくなるため，距離を測定する長い方の単語の長さで除して正規化する．単語 i，j 間のレーベンシュタイン距離を $L(i,j)$，単語 i の単語長を $\mathrm{length}(i)$ とすると，

$$d(i,j) = \frac{L(i,j)}{\max\{\mathrm{length}(i), \mathrm{length}(j)\}}$$

を正規化編集距離と呼び，オブジェクト間のメトリックとした．本稿では，単語データと呼ぶ．

3 つ目は，2014 年 4 月から 5 月までの Yahoo!ニュースの新聞記事データである．データベースから古い順に 5,000 記事を抽出した．含まれる単語数は 28,251 であり，記事間の類似度としてベクトル空間モデル [7] で頻繁に用いられる単語頻度ベクトル（Bag Of Words）間のコサイン類似度を採用する．記事 i，j 間のコサイン類似度 $s(i,j)$ を以

[*1] 平方根をとらない場合，三角不等式を満たさなくなるため，非類似度ではあるが距離ではなくなる．

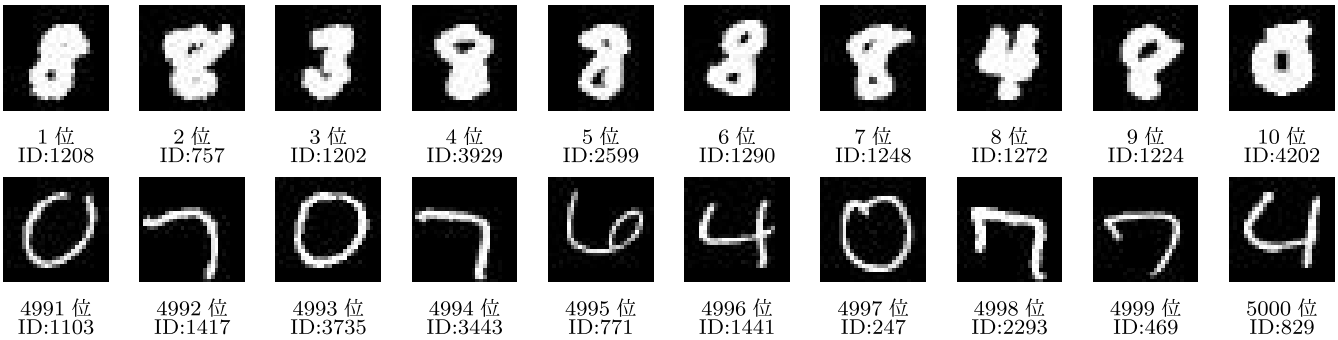


図 1 文字データ 混合中心性ランキング
Fig. 1 Mixedness centrality ranking of character data.

下のように距離に変換し，オブジェクト間のメトリックとした：

$$d(i,j) = \sqrt{2(1 - s(i,j))}.$$

距離の公理（三角不等式）を満たすように，一般的に用いられる平方根をとった距離を採用する．本稿では，文書データと呼ぶ．

4 つ目は，Content-based Photo Image Retrieval（CoPhIR）コレクションに含まれる Flickr の画像データである [1]．このデータベースでは画像データを MPEG-7 形式で保存しており，各画像はディスクリプタと呼ばれる特徴量で数値化されている．54,585,718 画像で構成されるデータベースからランダムに 5,000 画像を抽出して用いた．本稿では，Color Layout ディスクリプタに L_2 距離，Edge Histogram，Homogeneous Texture，Scalable Color，Color Structure ディスクリプタに L_1 距離を用いて，それぞれを所定の比で混合させたものをオブジェクト間のメトリックとした．本稿では，画像データと呼ぶ．

4.2 混合中心性ランキング結果

前述した 4 つのメトリック空間オブジェクトデータに対して，複数カテゴリに属するハイブリッドオブジェクトが抽出できるか否かについて，混合中心性ランキング上位・下位のオブジェクトを対象に考察する．

4.2.1 文字データ

文字データに対する混合中心性ランキングを**図 1** に示す．上位オブジェクトとして太い文字が多く抽出された．文字が太すぎると，他の数字との見分けがつかなくなる傾向にあるが，そのような誤識別率が高い文字が抽出されたといえる．**図 2** に，文字データの全オブジェクトを多次元尺度構成法により 2 次元に埋め込んだ結果を示す．図中の色の違いは，数字クラスの違いを表す．星で記されたオブジェクトは混合中心性上位 10 件のオブジェクトである．数字クラス ‘8’ を表すクラスタ（茶色）は，データ分布の中でも中心に位置するクラスタであることが分かる．図 2 より，定性的ではあるが，上位オブジェクトはクラスタ間

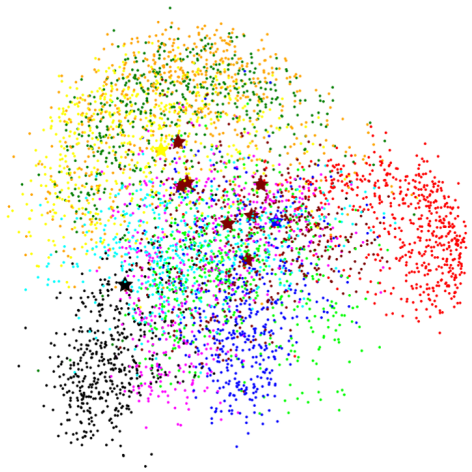


図 2 文字データの多次元尺度構成法による可視化結果
Fig. 2 Visualizatoin results of character data by MDS.

の狭間に存在するハイブリッド（この場合，誤識別率の高い）オブジェクトであることが分かる．このことを定量的に示す．オブジェクト u のクラスを $c(u)$，k 近傍までのオブジェクト集合を $\Gamma_k(u)$ とし，オブジェクト u のクラス一致率を以下のように定義する：

$$A_u(k) = \frac{|\{v : c(u) = c(v),\, v \in \Gamma_k(u)\}|}{k}.$$

図 3 に，混合中心性，凝集中心性における上位オブジェクト，および，K-medoids 法による 10 個の代表オブジェクトのクラス一致率の平均を示す．横軸は各上位オブジェクトから k 近傍のオブジェクト，縦軸は k 近傍オブジェクトのクラス一致率の平均を表す．赤●でプロットした混合中心性上位オブジェクトの平均クラス一致率は，青▲の凝集中心性上位オブジェクト，後述する K-medoids 法による代表オブジェクト（緑■）と比べて低いことが分かる．すなわち，各上位オブジェクトの近傍には，自身と異なるクラスのオブジェクトが存在しており，上位オブジェクトは複数のクラスタの狭間に存在することが定量的にも分かる．誤識別率が高いオブジェクトが重要ということではなく，複数のカテゴリの特徴量を併せ持つオブジェクトが抽出されることが確認できた．逆にランキング下位のオブジェクトを見ると細い文字が多く，上位オブジェクトとは

表 1 混合中心性ランキング
Table 1 Mixedness centrality rankings.

(a) 単語データ

順位	オブジェクト名
1	KITAKAMI
2	KITAYOSIWARA
3	SINANOSAKAI
4	SINANOKIZAKI
5	SIRAOI
6	HIGASIOOSAKA
7	SIMANOSITA
8	SAKAI
9	KASIHARA
10	KANISAWA
:	:
3254	SENCYOO
3255	EBICU
3256	HUZYUU
3257	OE
3258	YUU
3259	CUZU
3260	YUE
3261	ZYOONO
3262	ZEZE
3263	TENTOO

(b) 文書データ

順位	オブジェクト名
1	装着古い?次は埋め込み端末【経済・科学】
2	DV 離婚後に襲撃不安伝わらず【社会】
3	USJ「脱大阪」へ市と裁判も【政治・経済】
4	若田さんが会見【国際・科学・社会】
5	亀岡暴走 2 年声を上げ続ける【地域・社会】
6	男児 7 年放置?児相は迷子扱い【地域・政治】
7	ウルムチ 31 人死亡無差別テロ【国際・政治】
8	XP 使い続ける場合の対策 4 つ【科学・経済】
9	荷物検査はランダム AKB 会場【エンタメ・社会】
10	月給 20 万坪井智哉の米挑戦【スポーツ】
:	:
4991	中国側体当たり越の船体に穴【国際・政治】
4992	NFL 選手競馬で勝って札配る【スポーツ】
4993	総武線（快速）が運転再開【社会】
4994	脱「3K」目指す公衆トイレ【地域】
4995	総武線快速運転を再開【社会】
4996	ドイツ代表 DF，ロビーで放尿【スポーツ】
4997	東武東上線が運転再開【社会】
4998	岩佐が語る AKB と演歌の両立【エンタメ】
4999	中国機異常接近背景と意図【政治・国際】
5000	引退トワイライト EX の魅力【経済・政治】

(c) 画像データ

順位	オブジェクト概要
1	人物 3 人 (ID:334)
2	人物 2 人 (ID:3197)
3	人物 1 人＋動物 2 匹 (ID:3878)
4	人物 1 人 (ID:2131)
5	人物 2 人 (ID:3304)
6	人物 1 人 (ID:4775)
7	人物 2 人 (ID:615)
8	人物 3 人 (ID:3754)
9	動物 1 匹 (ID:4713)
10	人物 4 人 (ID:4614)
:	:
4991	夜景 (ID:2959)
4992	幾何学模様 (ID:3721)
4993	青空＋鳥 1 匹 (ID:2319)
4994	模様 (ID:315)
4995	暗い植物写真 (ID:1121)
4996	模様 (ID:1568)
4997	暗い海 (ID:538)
4998	林の中 (ID:1611)
4999	不明 (ID:1997)
5000	模様のような絵画 (ID:2093)

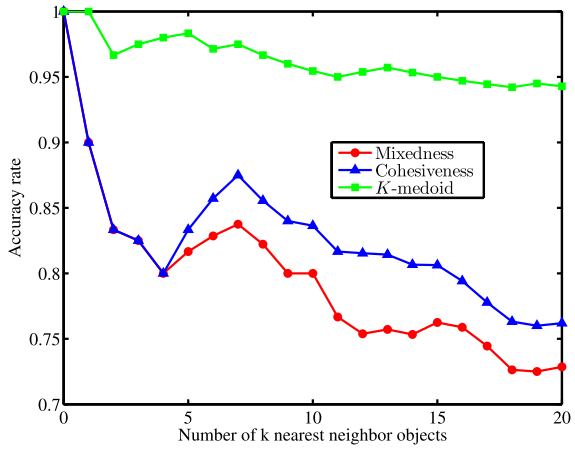


図 3 各種中心性上位オブジェクトに関するクラス一致率
Fig. 3 Agreement rate of classes between highly ranked object and its neighbors.

反対の性質を有するものが抽出されていた．

4.2.2 単語データ

単語データに対する混合中心性ランキングを**表 1** (a) に示す．上位オブジェクトとして，“KITAKAMI”，“KITAYOSIWARA”，“SINANOKIZAKI”，“HIGASIOOSAKA”，“KASIHARA”，“KANISAWA” などが抽出された．「北」，「西」，「南」，「北」などは駅名の先頭に，「町」，「山」，「原」，「崎」は末尾に現れることが多く，ある部分文字列を先頭に持つクラスタと，別の部分文字列を末尾に持つクラスタの狭間に存在する単語が上位オブジェクトとして抽出されたと考えられる．紙面の都合上すべては掲載できないが，11 位以降も上述の傾向に矛盾のない単語が多く抽出されている．すなわち，混合中心性の意図するオブジェクトが抽出されていた．

逆にランキング下位のオブジェクトを見ると，“SENCYOO”，“EBICU”，“HUZYUU”，“OE”，“ZEZE”，“TENTOO” などが抽出された．これらの単語は，“Z” や “Y” などの一般的に使用頻度の低いアルファベットを含む単語であるため，データ分布の端に存在し下位になったと考えられる．

4.2.3 文書データ

文書データに対する混合中心性ランキングを表 1 (b) に示す．表中【】の中は，ナイーブベイズにより学習したカテゴリの事後確率が高いカテゴリ名であり，複数存在する場合には「・」で区切って表す．上位オブジェクトは，経済と科学，政治と経済，国際と科学，エンタメと社会などの事後確率が高いカテゴリが複数存在する記事であり，各トピックが形成するクラスタの狭間に存在する記事が抽出されたと考えられる．すなわち，混合中心性の意図するオブジェクトが抽出されていた．

逆にランキング下位のオブジェクトを見ると，スポーツ，社会，地域，エンタメなどの単一のトピックが割り当てられた記事が名を連ねている．これらの記事は，事後確率がきわめてに高いカテゴリが 1 つだけ存在する記事であり，上位オブジェクトとは反対の性質を有するものと考え

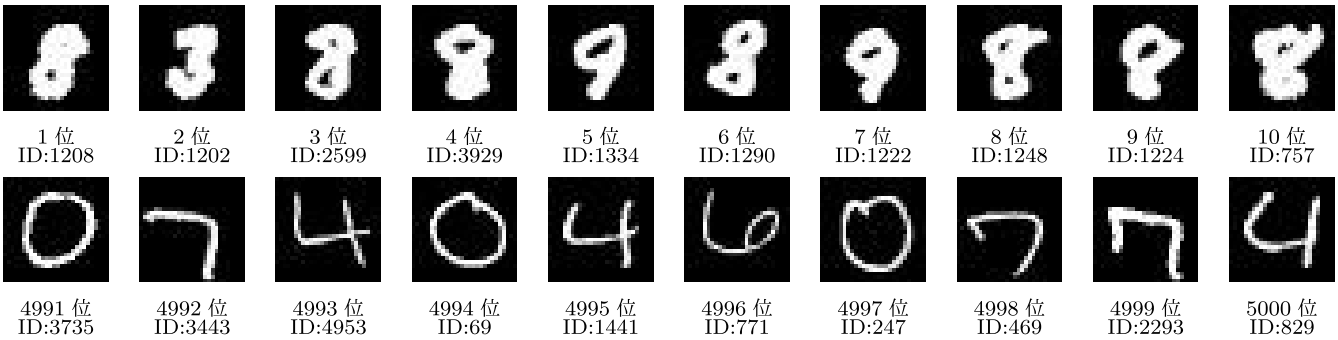


図 4 文字データ 凝集中心性ランキング
Fig. 4 Cohesiveness centrality ranking of character data.

られる．

4.2.4 画像データ

画像データに対する混合中心性ランキングを表 1 (c) に示す．なお，著作権に配慮して，Flickr の画像ではなく，画像の特徴を記述した．上位オブジェクトは，“複数の人物” や，“ペットと主人” の写真など，色彩豊かな画像が多く抽出された．なお，K-medoids 法などの代表的なクラスタリング手法により抽出されるクラスタが，色相で分けられる傾向にあることを確認済みである．すなわち，同一色相の画像がクラスタを形成している．上記のような色彩豊かな画像は，データ分布の中心や各クラスタ間の狭間に存在していると考えられ，混合中心性の意図するオブジェクトが抽出されることが確認できた．

逆にランキング下位のオブジェクトを見ると，“夜景” や “青空”，“暗い植物写真”，“暗い海”，“林の中” など，画面いっぱいに同じような色相，テクスチャが広がる写真が多く，クラスタの狭間というよりはクラスタの端に存在するようなオブジェクトといえ，上位オブジェクトとは反対の性質を有すると考えられる．

4.3 凝集中心性ランキング結果

類似オブジェクトが多数存在するオブジェクトが抽出できるか否かについて，凝集中心性ランキング上位・下位のオブジェクトを対象に考察する．

4.3.1 文字データ

文字データに対する凝集中心性ランキングを**図 4** に示す．上位オブジェクトは，混合中心性と同様に太い文字が多く抽出された．10 件中 8 件が同一文字である．文字が太い，すなわち，画素値が 0 でない画素が多く連なっていることになる．ここでは，画素値が 0 でない画素の数を有効次元数と呼ぶことにする．有効次元数が高いオブジェクトほど類似オブジェクトが多くなる傾向にあるため，近傍に多くの類似オブジェクトが存在するようなオブジェクトが抽出されたと考えられる．すなわち，凝集中心性の意図するオブジェクトが抽出された．逆にランキング下位のオブジェクトを見ると細い文字が多く，上位オブジェクトとは反対の性質を有するものが抽出された．

4.3.2 単語データ

単語データに対する凝集中心性ランキングを**表 2** (a) に示す．上位オブジェクトは，“KAMIKAWA”，“KAMIYAMA”，“KAMINAKA”，“KAMIHAMA”，“KASIHARA”，“KASIMADA” などが抽出された．本稿で使用した単語データの表記は，偶数番目の文字が母音である傾向が強いことを注意しておく．これらの単語は，“*A*I*A*A” というパターンの単語であり，同一パターンの単語どうしは非常に類似度が高く（距離が小さく），データ中で密に分布していることから，上位オブジェクトとして抽出されたと考えられる．紙面の都合上すべては表示できないが，11 位以降も上述の傾向に矛盾なく，上記の母音パターンあるいは他のパターンの単語群が多く抽出されている．すなわち，凝集中心性の意図するオブジェクトが抽出された．

逆にランキング下位のオブジェクトを見ると，“EBEOCU”，“HUCYUU”，“UCU”，“YUU”，“UBE” などが抽出された．これらの単語は，母音の配列が珍しく，類似するオブジェクトがデータ中にあまり存在しない単語なので，下位になったと考えられる．

4.3.3 文書データ

文書データに対する凝集中心性ランキングを表 2 (b) に示す．上位オブジェクトは，小保方晴子氏関連，田中将大選手関連，尖閣諸島問題関連などの記事が抽出された．これらの記事は，大きな話題や事件，問題に関するものであり，データ中に類似記事が多く，それらが密に分布していることから，上位オブジェクトとして抽出されたと考えられる．すなわち，凝集中心性の意図するオブジェクトが抽出された．

逆にランキング下位のオブジェクトを見ると，本願寺，能登，氷見などの明らかなローカルニュースが多い．これらは，記事内の文章が短く単語が少ない，類似記事がデータ中に含まれないなどの特徴を持ち，比較的小規模な記事であるため，上位オブジェクトとは反対の性質を有するものと考えられる．

表 2 凝集中心性ランキング
Table 2 Cohesiveness centrality rankings.

(a) 単語データ

順位	オブジェクト名
1	KAMIKAWA
2	KAMIYAMA
3	SIKAMA
4	KAMINAKA
5	KAMIHAMA
6	NAKAMA
7	KASIHARA
8	KASIMADA
9	KANISAWA
10	KASIMA
:	:
3254	PIQPU
3255	CYUUBU
3256	BECUDEN
3257	EHUE
3258	SEQPU
3259	EBECU
3260	NEQPU
3261	BEQPO
3262	ZEZE
3263	BEQPU

(b) 文書データ

順位	オブジェクト名
1	小保方氏の上司 責任どう説明【科学】
2	悪意明らか 小保方氏不正確定【科学】
3	日米会談「尖閣は安保対象」【国際・政治】
4	マー君 7 回 3 失点で開幕 4 連勝【スポーツ】
5	マー君 メジャー初登板で勝利【スポーツ】
6	ロシア軍事介入の懸念高まる【国際】
7	小保方氏の指導役 STAP は本物【科学】
8	笹井氏 STAP 研究停滞に懸念【科学】
9	小保方氏「清水の舞台から」【科学】
10	前回黒星マー君 1 失点で 7 勝目【スポーツ】
:	:
4991	どうなる 激減ブルートレイン【経済】
4992	岩佐が語る AKB と演歌の両立【エンタメ】
4993	動物写真家ネコの魅力【エンタメ・地域】
4994	本願寺派新門主就任 37 年ぶり【地域・社会】
4995	引退トワイライト EX の魅力【社会・経済】
4996	南明奈，ロリータ姿を披露【エンタメ】
4997	ダイオウイカ 能登でまた捕獲【地域】
4998	行司が不在 土俵入り遅れる【スポーツ】
4999	富山・氷見で大物 248 kg マグロ【地域】
5000	中尾明慶が整形疑惑を否定【エンタメ】

(c) 画像データ

順位	オブジェクト概要
1	人物 3 人 (ID:334)
2	人物 2 人 (ID:3197)
3	人物 2 人 (ID:615)
4	人物 3 人 (ID:2463)
5	人物 4 人 (ID:1888)
6	人物 2 人 (ID:3304)
7	人物 1 人＋動物 2 匹 (ID:3878)
8	室内のサボテン (ID:4723)
9	人物 1 人 (ID:2131)
10	人物 3 人 (ID:2171)
:	:
4991	カラフルな花畑 (ID:71)
4992	大量の観衆 (ID:1705)
4993	花畑 (ID:4813)
4994	車のバンパー (ID:1133)
4995	ビルのガラス (ID:453)
4996	模様 (ID:1568)
4997	林の中 (ID:1611)
4998	模様 (ID:315)
4999	不明 (ID:1997)
5000	模様のような絵画 (ID:2093)

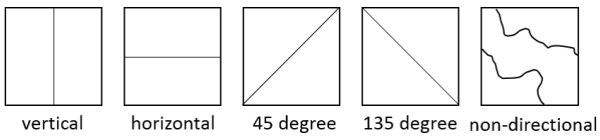


図 5 Edge histogram ディスクリプタ
Fig. 5 Discriptor of Edge histogram.

4.3.4 画像データ

画像データに対する凝集中心性ランキングを表 2 (c) に示す．上位オブジェクトに関しては，混合中心性と同様に，人物が被写体となっている画像が多く抽出された．これらの共通特徴として，被写体が画像全体に大きく映っている点があげられる．逆にランキング下位のオブジェクトを見ると，“花模様” や “大量の人物” など，曲線を多く含む画像が多くみられる．

上位の画像は被写体が人物であり，写真全体に被写体が大きく映っているため，人物の輪郭などが特徴的となる．あるいは，人物の背景にある物が特徴的となる場合がある．すなわち，**図 5** の 45 degree や 135 degree なエッジ，あるいは，vertical や horizontal なエッジが特徴的となる．これらの特徴を有する画像群が密集しているため，上位オブジェクトとして抽出されたと考えられる．また，下位の画像は花模様など輪郭が曲線状になっているものが多い．すなわち，図 5 の non-directional なエッジの特徴量が多い画像である．これらの特徴を有する画像は，曲線の種類が単純でなく多様であり，類似する画像が少なく，比較的疎に分布するオブジェクトが下位になっていると考えられる．これらより，凝集中心性の意図するランキング結果が得られることが確認できた．

4.4 両指標間の相関

上記の 4 つのメトリックデータに対する結果では，混合中心性と凝集性中心性によるランキングに相関がみられるデータが一部存在するため，上位オブジェクト集合の一致度について評価する．**図 6** は，横軸に順位 r を対数で，縦軸には以下に示す一致度をプロットした．

$$F(r) = \frac{2|M(r) \cap C(r)|}{|M(r)| + |C(r)|} = \frac{|M(r) \cap C(r)|}{r}$$

対数でプロットしたのは，上位ランキングの一致度をより強調するためである．ここで，$M(r)$ と $C(r)$ は，混合中心性，凝集中心性の上位 r 位までのオブジェクトの集合を表す．当然 r がオブジェクト数 N で一致度 $F(N) = 1$ となる[*2]．図 6 において，文字データ（Character）と画像データ（Image）に対する両中心性ランキングの一致度を見ると，全体的に比較的高い値を示しているのが分かる．すなわち，両指標により抽出されたオブジェクト群は比較的類似する傾向にある．一方，単語データ（Word）と文書データ（Document）では，全体的に比較的低い値を示し

[*2] 文字，文書，画像データで $N = 5{,}000$，単語データで $N = 3{,}263$ である．

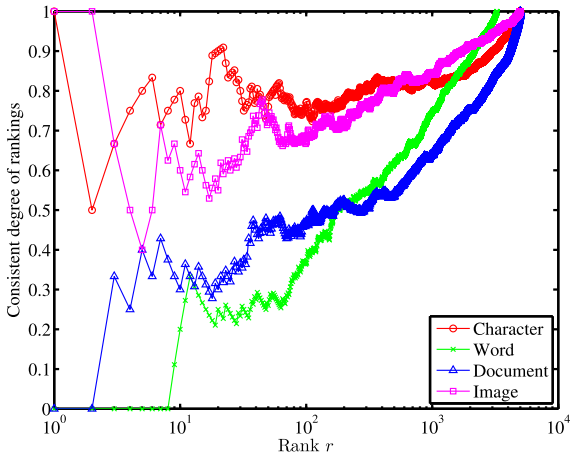


図 6 ランキング上位オブジェクト集合の一致度
Fig. 6 Agreement rate of highly ranked objects by each centrality measure.

ているのが分かる．すなわち，両指標により抽出されたオブジェクト群に，違いがあることが分かる．これらの違いは，データ分布の多峰性やクラスタ間の離れ具合などに影響されると考えられるが，詳細な分析は今後の課題である．

5. *K*-medoids 法による代表オブジェクトとの比較

メトリックデータの代表オブジェクト抽出や非階層クラスタリング手法として有名な K-medoids 法 [12] を比較手法として採用する．

5.1 *K*-medoids 法

K-medoids 法は，N 個のオブジェクト集合 U が与えられたとき，K 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（距離の小さい）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手法である．

K-medoids 法の解法には反復法や貪欲法があるが，解の一意性が保証される貪欲法に着目する．貪欲法は，代表オブジェクト集合を P とし，代表オブジェクトの候補オブジェクトを w とするとき，以下の目的関数を最小にするようなオブジェクト w を選び，代表オブジェクト集合 P を求める．

$$F(P \cup \{w\}) = \sum_{v \in U} \min\{D(v;P), d(v,w)\}. \qquad (6)$$

ここで，$D(v;P)$ は，すでに選定済みの代表オブジェクトとの距離の最小値を表し，$D(v;P) = \min_{p \in P}\{d(v,p)\}$ で定義される．この目的関数が有するサブモジュラ性により，貪欲解は厳密解ではないものの，最悪ケースの解品質が理論的に保証されていることから，ある程度妥当な精度で求めることができる [6]．本稿では，貪欲解法の各反復で得た解に対し局所改善を施すことにより，目的関数の改善を図った逐次局所改善付き貪欲法を採用する．

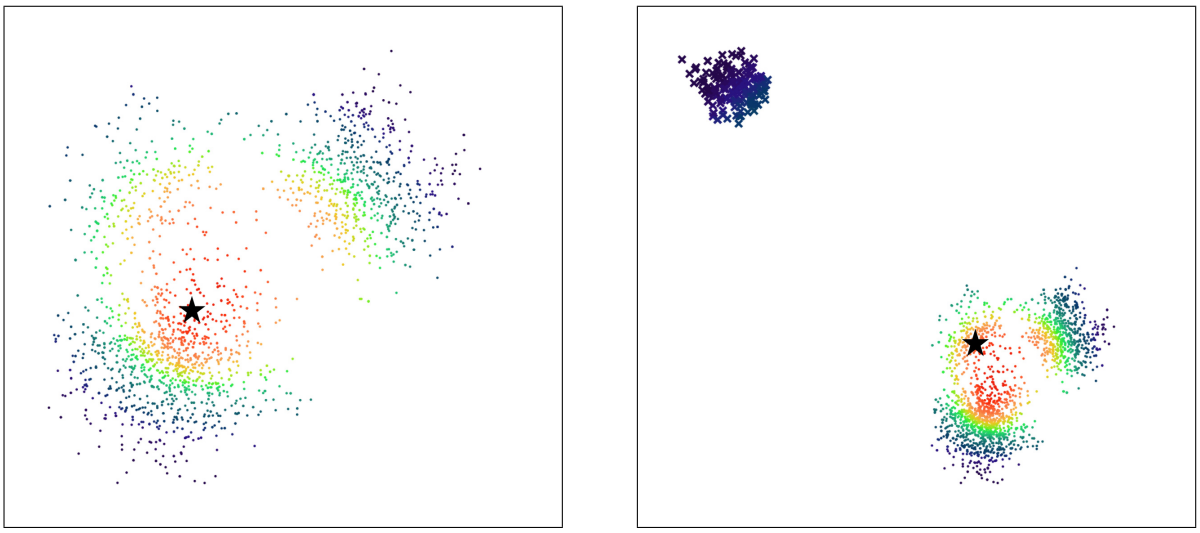
5.2 評価結果

$K = 10$ とした際の K-medoids 法により抽出した代表オブジェクトと提案指標の上位オブジェクトを比較して考察する．

K-medoids 法により，文字データでは，一部重複があるものの，“8”，“9”，“1”，“0”，“1”，“7”，“6”，“3”，“2”，“9” という異なるクラスタとして抽出された．単語データでは，“KASIMA”，“MINOO”，“OOKA”，“KAMIYAMA”，“HIGASISANO”，“SAKURA”，“TAMACI”，“KINOSAKI”，“SINAGAWA”，“SIZUMI” が抽出され，クラスタごとに “*A*I*A” 型，“*A*U*A” 型，“*I*U*I” 型のような固有のパターンが確認できた．文書データでは，“悪意明らか 小保方氏不正確定【科学】”，“マー君 7 回 3 失点で開幕 4 連勝【野球】”，“ロシア軍事介入の懸念高まる【国際】”，“ASKA 容疑者宅で薬物見つかる【社会】”，“沈没船 陰に隠れた家族の悲痛【国際】”，“集団自衛 行使容認に首相意欲【政治】”，“なでしこ 7 発快勝 新鋭が躍動【サッカー】”，“好業績相次ぐ 最終益 2.1 倍【経済】”，“PC 遠隔 4 事件への関与認める【社会】”，“AKB 握手会 川栄さん切られる【エンタメ】” が抽出され，科学，野球，国際，政治，経済，エンタメなど，クラスタごとに異なるカテゴリの記事が確認できた．画像データでは，“人物 1 人”，“複数の子供たち”，“暗闇に浮かぶデジタル時計”，“パーティー会場”，“仏像”，“白黒集合写真”，“グラス”，“閑静な住宅街”，“人物 2 人”，“橋” が抽出され，クラスタごとに異なる特徴の画像が確認できた．いずれのデータにおいても，抽出されたオブジェクト間に強い関連はみられず，各クラスタにおける代表的なオブジェクトが抽出できているといえる．これらの抽出結果は，凝集中心性の上位オブジェクトを幾分か含んでいる．K-medoids 法は，他のオブジェクトとの距離の和が最小であるオブジェクトを抽出する点で凝集中心性と同じ枠組みであるが，すでに選択したオブジェクトとの関係を考慮するか否かの点で異なる．また，混合中心性ではクラスタの狭間に存在するようなオブジェクトが抽出される傾向にあるが，K-medoids 法では各クラスタの中心に存在するようなオブジェクトが抽出される点で，反対の性質を有するといえる．すなわち，混合中心性の上位オブジェクトは複数カテゴリの性質を有するハイブリッドオブジェクトであり，K-medoids 法による代表オブジェクトは各クラスタの雛形オブジェクトであるといえ，どちらもテンプレートとして有用であることが示唆された．

6. 人工データによる評価

混合中心性により抽出されるオブジェクトの性質を明らかにするために，以下に示す人工データを用いて評価する．混合中心性の定義から，より多くのオブジェクトペア



(a) 外れ値なし　　(b) 外れ値あり（はずれ具合：20σ）

図 7 外れ値付き人工データにおける混合中心性ランキングと重心

Fig. 7 Mixedness centrality ranking and gravity center of synthetic data with outliers.

が生成する Lune に含まれるオブジェクトが上位オブジェクトとして抽出される．直感的には，オブジェクト分布の重心に位置するオブジェクトが該当すると考えられる．しかし，実際のオブジェクト群はメトリック空間，あるいは高次元ベクトル空間の点として表現され，必ずしも分布の重心オブジェクトが抽出されるとは限らない．

以上のことを検証するために，レプリカ交換モンテカルロ法により生成した 2 次元混合ガウス分布に外れ値を混入させた人工データを用いる．オブジェクト間の距離には L_2 距離を用いる．**図 7** (a) は，オブジェクト数 2,300 のオブジェクトからなる 3 峰のガウス混合分布であり，外れ値を混入させる前の人工データである．各ガウス分布のオブジェクト数は，1,100，900，300 であり，標準偏差はすべて $\sigma = 1$ である．図 7 (b) は，図 7 (a) の重心から距離が 20σ だけ離れた位置を中心に，200 個の外れ値オブジェクトを混入させた人工データである．各図において，右下に位置するオリジナルオブジェクトは○，左上に位置する外れ値オブジェクトは×，全オブジェクトの重心を黒い★でプロットした．また，混合中心性ランキング上位オブジェクトは赤，下位は青のグラデーションで彩色した．

図 7 (a) では，混合中心性上位オブジェクトとして 3 つのガウス分布の狭間，すなわち，各クラスタの狭間にあるオブジェクトが抽出されていることが分かる．重心もおおよそ上位オブジェクトと同様の位置に存在する．図 7 (b) では，外れ値に引っ張られ重心がずれていることが分かる．一方混合中心性上位オブジェクトの位置はあまり変化しておらず，クラスタ間の狭間のオブジェクトを頑健に抽出していることが分かる．図示はしていないが，外れ値オブジェクト群をさらに離すことで重心はさらにずれるが，混合中心性上位オブジェクトの位置は大きく変化しないことが確認できた．

たとえば文書データなどでは，各記事は単語頻度ベクトルで表現される．同一トピックのオブジェクトペア距離はある程度近いものの，異なるトピックのオブジェクトペア距離は，高次元ベクトルで両方が値を持つ要素が限られるので，非常に大きくなる．上述の人工データは，このような状況を反映したものである．すなわち，記事群の中に特異なキーワードを有する記事群（理研の小保方晴子氏関連，楽天の田中将大選手関連）が外れ値のような役割をしていると考えられる．このような実データにおいて，単純な重心では外れ値に引っ張られるのに対し，ペアワイズで計算される混合中心性は頑健であり，クラスタ間のオブジェクトを抽出できたと考えられる．

7. おわりに

本稿では，ネットワークから媒介性と近接性の高いノードを抽出する指標である媒介中心性，近接中心性をベースに，メトリック空間オブジェクトから混合性と凝集性の高いオブジェクトを抽出する指標を提案した．複数のクラスタの狭間に存在するハイブリッドなオブジェクトを抽出する混合中心性，類似オブジェクトが多く存在するオブジェクトを抽出する凝集中心性を提案し，両指標により抽出されるオブジェクトについて考察し，提案指標を評価した．4 つのメトリック空間データを用いた評価実験より，提案中心性は，凝集性の高い，あるいは，混合性の高いオブジェクトを抽出可能であることを示した．さらに，K-medoids 法による代表オブジェクトとの比較においては，異なる性質ではあるものの，データ分布を代表する雛形となりうるオブジェクトを抽出できることが示唆された．また，外れ値を含む人工データによる評価では，重心のように外れ値に影響を受けやすい指標とは異なり，クラスタ間の狭間に存在するオブジェクトを頑健に抽出できることが確認で

きた．

今後は，確率分布やグラフ構造など，多様なメトリック空間オブジェクトを対象に評価し，提案指標の有効性を確認していくつもりである．さらに，抽出されたオブジェクトとクラスタの関係を定量的に評価していきたい．

謝辞 本研究は，JSPS 科研費（No.15J00735），（No.16K16154）の助成を受けたものである．

参考文献

[1] Bolettieri, P., Esuli, A., Falchi, F., Lucchese, C., Perego, R., Piccioli, T. and Rabitti, F.: CoPhIR: a Test Collection for Content-Based Image Retrieval, *CoRR*, Vol.abs/0905.4627v2 (2009).

[2] Brandes, U.: A Faster Algorithm for Betweenness Centrality, *Journal of Mathematical Sociology*, Vol.25, pp.163–177 (2001).

[3] Freeman, L.: Centrality in social networks: Conceptual clarification, *Social Networks*, Vol.1, No.3, pp.215–239 (online), DOI: 10.1016/0378-8733(78)90021-7 (1979).

[4] Lee, J.A. and Verleysen, M.: *Nonlinear dimensionality reduction*, Springer, New York; London (2007).

[5] Levenshtein, V.: Binary Codes Capable of Correcting Deletions, Insertions and Reversals, *Soviet Physics Doklady*, Vol.10, p.707 (1966).

[6] Nemhauser, G.L., Wolsey, L.A. and Fisher, M.L.: An Analysis of Approximations for Maximizing Submodular Set Functions, *Mathematical Programming*, Vol.14, pp.265–294 (1978).

[7] Salton, G., Wong, A. and Yang, C.S.: A Vector Space Model for Automatic Indexing, *Comm. ACM*, Vol.18, No.11, pp.613–620 (1975).

[8] Seewald, A.K.: On the brittleness of handwritten digit recognition models, *ISRN Machine Vision*, Vol.2012 (2011).

[9] Supowit, K.J.: The Relative Neighborhood Graph, with an Application to Minimum Spanning Trees, *J. ACM*, Vol.30, No.3, pp.428–448 (online), DOI: 10.1145/2402.322386 (1983).

[10] Tenenbaum, J.B., Silva, V. and Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction, *Science*, Vol.290, No.5500, pp.2319–2323 (2000).

[11] Torgerson, W.: Multidimensional scaling: I. Theory and method, *Psychometrika*, Vol.17, pp.401–419 (1952).

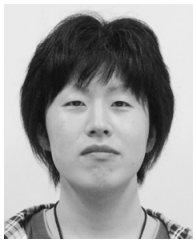
[12] Vinod, H.: *Integer Programming and The Theory of Grouping*, Vol.64, An Official Journal of the American Statistical Association (1969).

[13] Wasserman, S. and Faust, K.: *Social Network Analysis: Methods and Applications*, Cambridge University Press (1994).

[14] 牧野正三，二矢田勝行，真船裕雄，城戸健一：東北大–松下単語音声データベース，日本音響学会誌小特集，Vol.48, No.12, pp.899–905 (1992).

[15] 伏見卓恭，斉藤和巳，武藤伸明，池田哲夫，風間一洋：道路ネットワークに対する実距離を用いた中心性指標の提案と応用，ネットワークが創発する知能研究会（JWEIN2014）(2014).

[16] 伏見卓恭，斉藤和巳，池田哲夫，武藤伸明：ノード群の協調的振舞いに着目した集合媒介中心性の提案と応用，電子情報通信学会和文論文誌 D，Vol.J96-D, No.5, pp.1158–1165 (2013).



伏見 卓恭 （正会員）

2011 年静岡県立大学大学院経営情報学研究科修士課程修了．2014 年静岡県立大学大学院経営情報イノベーション研究科博士後期課程修了．同年静岡県立大学大学院経営情報学部客員研究員．2015 年筑波大学図書館情報メディア系特別研究員（PD）．2017 年より東京工科大学コンピュータサイエンス学部助教．複雑ネットワーク，可視化の研究に従事．博士（学術）．人工知能学会，日本データベース学会各会員．

斉藤 和巳 （正会員）

1985 年慶應義塾大学理工学部数理科学科数学専攻卒業．1998 年東京大学博士（工学）．2007 年より静岡県立大学経営情報学部教授．複雑ネットワークの研究に従事．電子情報通信学会，人工知能学会，日本神経回路学会，日本応用数理学会，日本行動計量学会，日本データベース学会各会員．

風間 一洋 （正会員）

1988 年京都大学大学院工学研究科精密工学専攻修士課程修了．同年日本電信電話（株）入社．2005 年京都大学大学院情報学研究科システム科学専攻博士課程修了．2012 年より和歌山大学システム工学部教授．Web 情報検索，Web マイニングの研究に従事．博士（情報学）．人工知能学会，日本ソフトウェア科学会，日本データベース学会，ACM 各会員．