RF-001

メトリック空間オブジェクトを対象とした中心性指標の提案
Proposing centrality measures intended for metric object data

伏見 卓恭 † 斉藤 和巳 ‡ 風間 一洋 §
Takayasu Fushimi Kazumi Saito Kazuhiro Kazama

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一層重要になっている．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 [1]．さらに，データに含まれるオブジェク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るなど注目を集めている [2]．

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て，ベクトルで表現できないオブジェクト群に対して，その中から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，オブジェクトを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，オブジェクト間の距離が与えら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．

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 [3]．重要ノードの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 [4]．

本研究では，ネットワーク中心性指標をメトリック空間オブジェクトに適用できるように拡張した指標を提案する．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の総和が最小であるオブジェクトが，オブジェクト集合の中で中心的な位置に存在すると想定でき，これを抽出することが最もシンプルな方法である．本稿では，このシンプルな方法をメトリック空間オブジェクトに対する近接中心性として定義する．しかし，このシンプルな方法では，データに複数のクラスタが存在するような状況のとき，最も中心に位置するクラスタに含まれるオブジェクトばかりが抽出されてしまう．この点を克服するため，他のオブジェクト間の媒介度という視点において中心的な存在を抽出する媒介中心性も提案する．

本稿の構成は以下の通りである．2 章でネットワークを対象とした既存の媒介，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媒介，近接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考察し，評価する．5 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 [3] の説明に準じて説明する．ノード集合 V，リンク集合 E からなる無向ネットワークを $G = (V, E)$ と表記する．文献 [3] と同様に，ネットワークとしては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

ノード v の媒介度とは，任意のノードペアを結ぶパスを，どの程度媒介しているかを示す指標である．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観に基づいた中心性であり，任意のノードペア間の最短パスのうち，媒介しているパスの割合によりノードをランキングするものである．ノード v の媒介中心性 $\mathrm{bwc}(v)$ を以下のように定義する．

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

ここで，$\sigma_{s,t}$ は始点ノード s と 終点ノード t 間の最短パス数であり，$\sigma_{s,t}(v)$ はノード v を通るノード s,t 間の最短パス数を表す．媒介中心性は，始点と終点ノードをランダムに選んだ際，ノード v が最短パス上に存在する確率を表していると言い換えられる．

2.2. 近接中心性

ノード v の近接度とは，ネットワーク内の他のノードへの近さを表す指標である．近接中心性とは，他の多くのノードへ少ないステップで行ける，ネットワークの中心にいるようなノードは重要であるという直観に基づいた中心性指標であり，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ものである．ノード v の近接中心性 $\mathrm{clc}(v)$ を以下のように定義する．

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

ここで $g(v,u)$ は，ノード v とノード u 間の最短パス長である．

†静岡県立大学大学院経営情報イノベーション研究科，日本学術振興会特別研究員 (PD)
‡静岡県立大学大学院経営情報イノベーション研究科
§和歌山大学システム工学部

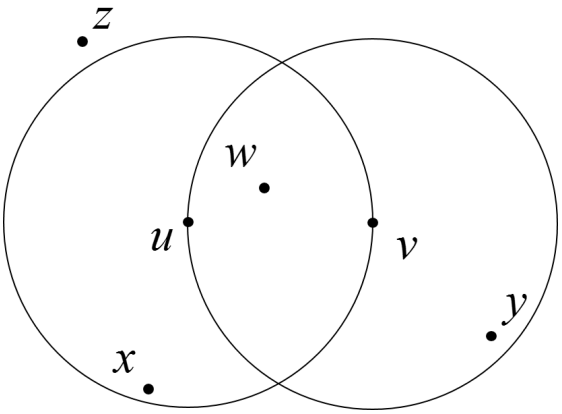


図 1: 媒介度と Lune

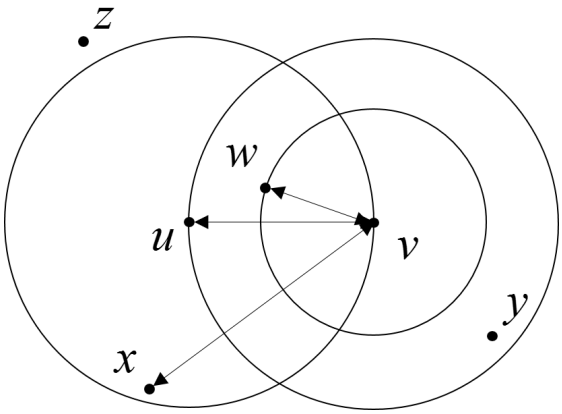


図 2: Lune に含まれるか否かの判定

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媒介中心性，近接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. メトリック媒介中心性

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．これを基にして，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媒介度を拡張定義する．オブジェクトペア間に存在するか否かについて，相対近傍グラフ [5] 構築における Lune の概念を用いる．

オブジェクト u と v の間に w が存在するか否かは，図 1 のように，u と v を中心とした半径 $d(u,v)$ の円の交わり部分（Lune）に含まれるか否かにより判定する．図 1 の例では，オブジェクト w のみが u, v 間に存在し，w の媒介度として数え上げる．

形式的には，オブジェクト u, v を中心とした半径 $d(u,v)$ の円でできる Lune を $L(u,v)$ と表記する．オブジェクト w が u, v 間に含まれるか否かを，

$$\delta_{u,v}(w) = \begin{cases} 1 & \text{if} \quad w \in L(u,v) \\ 0 & \text{otherwise} \end{cases}$$

とし，オブジェクト w の媒介度を以下のように定義する．

$$\mathrm{mbwc}(w) = \frac{\sum_{u \in U \setminus \{w\}} \sum_{v \in U \setminus \{u,w\}} \delta_{u,v}(w)}{(|U|-1)(|U|-2)} \qquad (3)$$

メトリック空間においては，任意のオブジェクト u から距離の近い順に w と $d(u,w) < d(u,v)$ となる v を比較し，$d(w,v) \leq d(u,v)$ の場合は $w \in L(u,v)$ と判断できる（図 2 参照）．図 2 のように，オブジェクト x は $d(x,v) > d(u,v)$ であるため，$x \notin L(u,v)$ と判断できる．

全てのノードを $\mathrm{mbwc}(w)$ の値で降順ソートし，ランキング上位のノードをメトリック空間における媒介中心オブジェクトとして抽出する．これにより，オブジェクト群が密集している部分空間において中心に位置するオブジェクトを抽出する．

3.2. メトリック近接中心性

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．これを基にして，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近接度を拡張定義する．

オブジェクト u と w 間の距離 $d(u,w)$ に対し，オブジェクト w の近接度を以下のように定義する．

$$\mathrm{mclc}(w) = \frac{\sum_{u \in U \setminus \{w\}} d(u,w)^{-1}}{(|U|-1)} \qquad (4)$$

全てのノードを $\mathrm{mclc}(w)$ の値で降順ソートし，ランキング上位のノードをメトリック空間における近接中心オブジェクトとして抽出する．これにより，他のオブジェクトへの平均距離が小さい，空間の中心に位置するオブジェクトを抽出する．

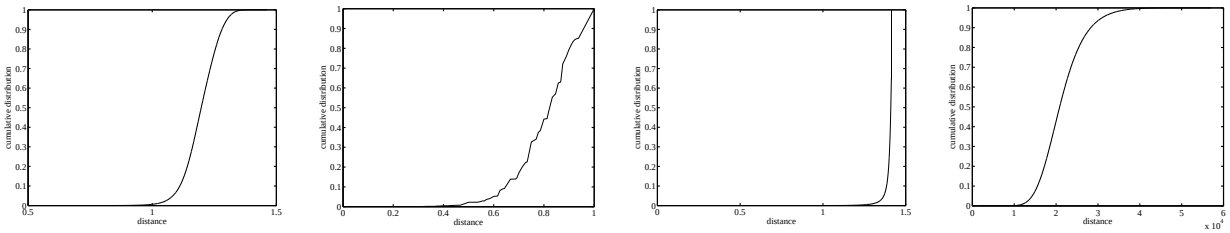
4. 評価実験

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 つ目のデータは，2000 年 1 月 4 日から 2009 年 12 月 30 日までの 10 年間の 2,457 営業日間において，東証一部に上場している 830 銘柄の株価終値のデータである．株価銘柄間の類似度として，Mantegna の文献 [6] にあるように，株価対数リターンの相関係数を用いる．株価対数リターンとは，連続する 2 つの時刻における株価の比率（変化率）に対数を掛けたものである．銘柄 i の t 日目と $t+1$ 日目の終値を $x_{i,t}$ と $x_{i,t+1}$ とすると，$y_{i,t} = \log(x_{i,t+1}) - \log(x_{i,t})$ が株価対数リターンである．銘柄 i, j 間の相関係数 $r(i,j)$ を

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



(a) 株価データ　(b) 文字データ　(c) 文書データ　(d) 画像データ

図 3: 評価データの累積距離分布

と変換した距離をオブジェクト間のメトリックとした．距離の累積分布を図 3(a) に示す．本稿では，株価データと呼ぶ．

2 つ目のデータは,「東北大・松下単語音声データベース Vol.5」に含まれるデータを抽出したものである．このデータベースに含まれる 3,263 個の日本の駅名をローマ字表記した文字データである．文字列間の類似度として，編集距離の代表例であるレーベンシュタイン距離 [7] を用いる．実際には文字 (character) の挿入・削除・置換を同コストとした最小の操作回数である．ただし，そのままでは長い文字列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文字列の長さで除して正規化する．文字列 i, j 間のレーベンシュタイン距離を $L(i, j)$，文字列 i の単語長を length(i) とすると，

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

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．距離の累積分布を図 3(b) に示す．本稿では，文字データと呼ぶ．

3 つ目のデータは，1992 年から 2002 年までの毎日新聞国際面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 51,030 であり，記事間の類似度として単語頻度ベクトル（Bag Of Words）間のコサイン類似度を用いる．記事 i, j 間のコサイン類似度 $s(i, j)$ を

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

と変換した距離をオブジェクト間のメトリックとした．距離の累積分布を図 3(c) に示す．本稿では，文書データと呼ぶ．

4 つ目のデータは，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 [8]．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 画像を抽出した．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．本稿では，Color Layout ディスクリプタに L2 距離，Edge Histogram，Homogeneous Texture，Scalable Color，Color Structure ディスクリプタに L1 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．距離の累積分布を図 3(d) に示す．本稿では，画像データと呼ぶ．

4.2. 媒介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，媒介中心性により抽出されたランキング上位オブジェクトについて考察する．

株価データに対する媒介中心性ランキングを表 1 に示す．上位オブジェクトとして,“ 岡三証券グループ ”,“ 東海東京フィナンシャル・ホールディングス ”,“ 東洋証券 ”,“ みずほインベスターズ証券 ”,“ みずほ証券 ”などの証券業の銘柄や“ 住友商事 ”,“ 三菱電機 ”,“ 豊田通商 ”,“ 豊田自動織機 ”,“ 東レ ”,“ 旭化成 ”などの大手企業の銘柄が多く抽出された．これらの銘柄は他の銘柄への影響力が強く，類似する（距離の小さい）オブジェクトが周辺に存在するため，上位オブジェクトとして抽出されたと考えられる．

逆にランキング下位のオブジェクトを見ると，期間中に株式分割を行うなど，それぞれ特異な変動をする銘柄が多く，類似オブジェクトが少ない，いわゆる外れ値のようなオブジェクトが抽出された．

文字データに対する媒介中心性ランキングを表 2 に示す．上位オブジェクトとして,“ KITAKAMI ”,“ KITAYOSIWARA ”,“ SINANOSAKAI ”,“ SINANOKIZAKI ”,“ HIGASIOOSAKA ”,“ MINAMIOOMACI ”などの駅名が多く抽出された．これらの駅名は,「東西南北」や「〜町」,「〜山」といった接頭辞，接尾辞を含む駅名であるため，類似する（距離の小さい）オブジェクトが周辺に存在するため，上位オブジェクトとして抽出されたと考えられる．

逆にランキング下位のオブジェクトを見ると,“ BEQPU ”,“ SENCYOO ”,“ EBICU ”,“ HUZYUU ”,“ OE ”,“ ZEZE ”,“ TENTOO ”など，特殊な名前の駅名が多く，類似オブジェクトが少ない駅名がランクインしていた．

文書データに対する媒介中心性ランキングを表 3 に示す．上位オブジェクトとして,“ ロシア情勢 ”に関連する記事,“ カンボジアの選挙 ”に関する記事,“ ボスニアの独立 ”に関する記事などが多く抽出された．このデータは 1992 年からの国際面記事であるため，90 年代前半の世界における大きな出来事（ソ連崩壊後のロシア情勢，国際連合カンボジア暫定統治機構によるカンボジア統治など）の記事が多数発表され，これらに類似する記事が周辺に存在するため，上位オブジェクトとして抽出されたと考えられる．

逆にランキング下位のオブジェクトを見ると，国際面の一面に掲載される記事とは異なる内容のものが多

い．また，国内に閉じた内容の記事も多く含まれ，比較的他の記事との類似度が低いためと考えられる．

画像データに対する媒介中心性ランキングを表 4 に示す．なお，flickr の画像を掲載する代わりに，その特徴について記述する．上位オブジェクトとして,“ 複数の人物 ”や,“ ペットと主人 ”の写真などが多く抽出された．flickr や Facebook などの写真共有サイトでは，ユーザにより撮影されたユーザやその友人の写真が多く投稿，共有されるケースが多い．従って，データ中には人物が被写体となった写真が中心に分布しており，人物写真に類似する写真が周辺に存在するため，上位オブジェクトとして抽出されたと考えられる．

逆にランキング下位のオブジェクトを見ると,“ 夜景 ”や“ 模様 ”など，画面いっぱいに同じようなテクスチャが広がる写真が多く，類似オブジェクトが少ない，いわゆる外れ値のようなオブジェクトが抽出された．

このように，提案した媒介中心性により，メトリック空間オブジェクトの中から，ある程度妥当な，中心となる重要オブジェクトを抽出できたことが示された．

4.3. 媒介中心性と近接中心性の相違点

媒介中心性，近接中心性それぞれにより抽出されたオブジェクトの違いを示し，提案した媒介中心性の有用性を示す．

図 4 に両手法によるランキング結果の一致度を示す．横軸に順位 r，縦軸には以下に示す一致度をプロットした．

$$F(r) = \frac{2 \cdot |B(r) \cap C(r)|}{|B(r)| + |C(R)|} = \frac{|B(r) \cap C(r)|}{|B(r)|}$$

ここで，$B(r)$ と $C(r)$ は，媒介中心性，近接中心性の上位 r 位までのオブジェクトの集合を表す．

図 4(a) の株価データ，および，図 4(d) の画像データに対する両指標の一致度を見ると，どの順位までででも比較的高い値を示しているのがわかる．すなわち，両指標により抽出されたオブジェクト群には大きな違いがないことがわかる．このことから，データに含まれるオブジェクト全体が一つの大きな群を成していることが推定できる．

一方，図 4(b) の文字データ，および，図 4(c) の文書データに対する両指標の一致度を見ると，どの順位までででも比較的低い値を示しているのがわかる．すなわち，両指標により抽出されたオブジェクト群に，大きな違いがあることがわかる.「はじめに」で述べたように，データの中に複数のクラスタのようなものが存在する場合，定義より近接中心性では全オブジェクトの中心に存在するオブジェクトしか抽出されない．媒介中心性では，複数のクラスタが存在している場合においても，自身の周辺に類似オブジェクトが多く存在する場合，それらオブジェクトペアがなす Lune に含まれる回数が多くなるため，上位オブジェクトとして抽出されると考えられる．実際に単語データでは，近接中心性の上位オブジェクトとして,“ KA ”,“ MA ”,“ SI ”,“ MI ”などの文字を含む駅名ばかりが抽出された．文書データでは，近接中心性の上位オブジェクトとして,“ ボスニア ”,“ セルビア ”を多く含む記事ばかりが抽出された．媒介中心性では,“ ロシア ”,“ カンボジア ”など各トピックの主要ニュース記事が抽出されたことと比較すると，全体の中心のみを抽出していることがうかがえる．

このように，提案した媒介中心性により，比較的単純かつ直感的な近接中心性とは異なり，メトリック空間におけるオブジェクトの分布が単一でない場合でも，ある程度妥当な重要オブジェクトが抽出できたことが示された．

4.4. 媒介中心性による重要オブジェクトの性質

ここでは，媒介中心性と近接中心性により抽出されたオブジェクトに顕著な違いのあった単語データと文書データの結果について考察する．

まず単語データでは,「東西南北」や「〜町」,「〜山」といった接頭辞，接尾辞を含む駅名が多く抽出された．これらオブジェクト群の近傍（距離が小さい）オブジェクトをみると,「東」や「北」という接頭辞を含むオブジェクトばかりが存在し，これらからなるクラスタの「中心」に近い位置にあるオブジェクトが抽出されたことが示唆される．一方，接頭辞も接尾辞も含むような重要オブジェクトの近傍をみると，類似する接頭辞，または，接尾辞を含むオブジェクト群が混在しており，ある接頭辞クラスタと別の接尾辞クラスタの「狭間」に位置するオブジェクトが抽出されたことが示唆される．

次に文書データでは，前述したように“ ロシア ”,“ カンボジア ”など各トピックの記事が抽出されているが，近傍オブジェクトをみると，重要オブジェクトと同様のトピックを持つオブジェクトが多く存在した．さらに複数のトピックを併せ持つようなオブジェクトも存在することから，各トピッククラスタの「中心」やクラスタ間の「狭間」に位置するオブジェクトが抽出されたことが示唆される．

5. おわりに

本稿では，ネットワークから重要ノードを抽出する指標の代表例である媒介中心性，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中心となるような重要オブジェクトを抽出する指標を提案した．他のオブジェクトとの距離が小さいオブジェクトを抽出する近接中心性と他のオブジェクト間に存在する度合が大きいオブジェクトを抽出する媒介中心性を提案し，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，提案媒介中心性はある程度妥当な重要オブジェクトを抽出可能であることを示した．また，単純な近接中心性と比較して，媒介中心性では空間内の分布に偏りがある場合でも妥当な結果が得られることが示唆された．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 本研究は科学研究費補助金 (No.25・10411) の補助を受けた．

表 1: 株価データ　媒介中心性ランキング

順位	オブジェクト名
1	岡三証券グループ
2	東海東京フィナンシャル H
3	東洋証券
4	みずほインベスターズ証券
5	みずほ証券
6	丸三証券
7	住友商事
8	洋紡
9	バンドー化学
10	三ツ星ベルト
：	：
821	ファンケル
822	パーク２４
823	サクラダ
824	モリテックス
825	サイゼリヤ
826	エコナック H
827	田崎真珠
828	山水電気
829	ＪＴ
830	東宝

表 2: 文字データ　媒介中心性ランキング

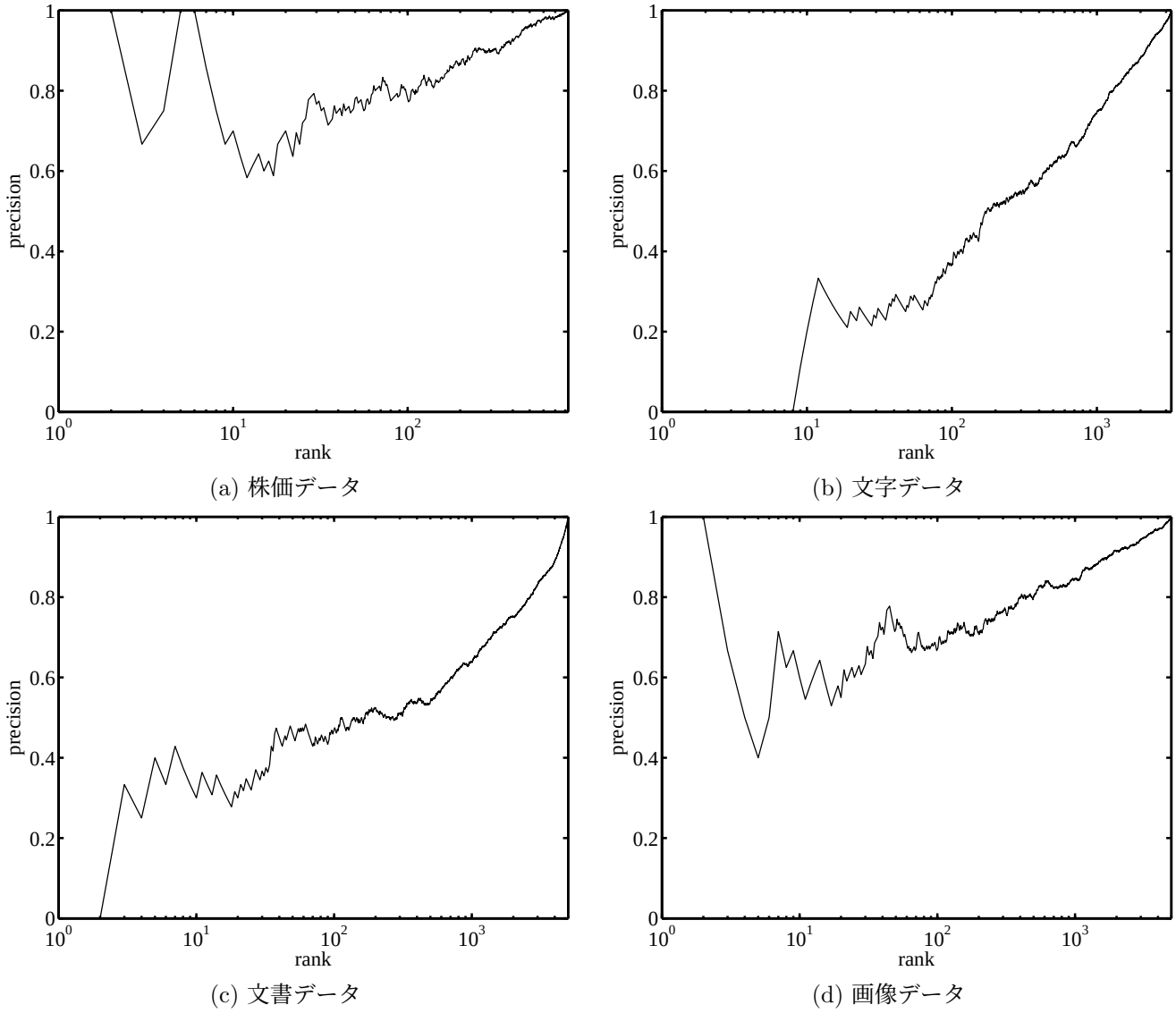
順位	オブジェクト名
1	KITAKAMI
2	KITAYOSIWARA
3	SINANOSAKAI
4	SINANOKIZAKI
5	SIRAOI
6	HIGASIOOSAKA
7	SIMANOSITA
8	SAKAI
9	KASIHARA
10	KANISAWA
：	：
3254	SENCYOO
3255	EBICU
3256	HUZYUU
3257	OE
3258	YUU
3259	CUZU
3260	YUE
3261	ZYOONO
3262	ZEZE
3263	TENTOO

表 3: 文書データ　媒介中心性ランキング

順位	オブジェクト名
1	露政治危機、経済改革の停滞が拍車
2	カンボジア総選挙
3	ロシアと中国の思惑 政治力低下恐れる
4	ボスニア和平はいつ… 米欧、足並みに乱れ
5	亀裂の兆しを見せ始めたフィリピン共産党
6	新段階迎えた米露関係
7	北朝鮮のＮＰＴ復帰は？ 米朝会談に注目
8	露大統領の国民投票勝利 国民不満あるが逆戻り拒む
9	全人代後の中国 香港民主化 原則堅持か経済利益か
10	米政府手詰まり状態　ボスニア問題
：	：
4991	中国のハイジャック
4992	中東和平交渉
4993	アゼルバイジャン領フィズリ占拠
4994	韓国、日本の歌謡曲公演を初許可
4995	南アフリカで米国の白人女子留学生刺殺
4996	ＮＹタイムズ、ボストン・グローブ紙買収
4997	ドイツ極右がナチスのヘス副総統追悼デモ
4998	モスタルに２回目の救援物資投下
4999	南アフリカ企業「濃縮ウラン売却を計画」
5000	イランで７０億バレルの油田発見

表 4: 画像データ　媒介中心性ランキング

順位	オブジェクト概要
1	人物 3 人
2	人物 2 人
3	人物 1 人＋動物 2 匹
4	人物 1 人
5	人物 2 人
6	人物 1 人
7	人物 2 人
8	人物 3 人
9	動物 1 匹
10	人物 4 人
：	：
4991	夜景
4992	幾何模様
4993	青空＋鳥 1 匹
4994	模様
4995	暗い植物写真
4996	模様
4997	暗い海
4998	林の中
4999	不明
5000	模様のような絵画



(a) 株価データ　(b) 文字データ

(c) 文書データ　(d) 画像データ

図 4: 近接中心性と媒介中心性

参考文献

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

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

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

[4] 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 (2013-05).

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

[6] Mantegna, R. N.: Hierarchical structure in financial markets, *European Physical Journal B*, Vol. 11, pp. 193–197 (1999).

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

[8] 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).