様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ９ 月 １３ 日現在

機関番号：５２６０５
研究種目：基盤研究(C)（一般）
研究期間：2016～2019
課題番号：１６Ｋ００１６５

研究課題名（和文）相互接続に基づいたクラスタリング手法の開発

研究課題名（英文）A Clustering Algorithm based on Mutually Ranking

研究代表者
小早川　倫広（Michihiro, Kobayakawa）

東京都立産業技術高等専門学校・ものづくり工学科・教授

研究者番号：００３３４５８２

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：データ群を分類する場合、各データから特徴量ベクトルを算出し、算出した特徴ベクトル間の距離が用いられる。しかし、データから特徴ベクトルを抽出できるとは限らない。この場合、距離を使用することはできない。この場合、距離の使用を前提としないクラスタリングアルゴリズムが必須となる。本研究では、相互に類似しているデータ間では、検索を実施した場合、相互に検索結果が上位になるという事象を、相互隣接グラフ（MNN-Graph）表現し、相互隣接グラフのクリーク発見するというものである。提案アルゴリズムは、閾値のチューニング・速度等に問題があるが、クラスタリング過程が説明可能であることが分かった。

研究成果の学術的意義や社会的意義
Society5.0を牽引するコア技術として、データ分析技術が必須となる。現在、AI等を活用したデータ分析技術が盛んに開発されている。しかし、データ分析における特徴量の選定により、分析すらできないことがある。本研究は、データ同士が相互に類似しているというシンプルな特徴を用いたクラスタリングアルゴリズムであり、クラスタリング生成の構造がシンプルである。したがって、データ間の類似の尺度が距離の公理を満たす・満たさないに関わらず適用可能である。精度・速度等が不足していることはあるが、汎用なクラスタリングアルゴリズムとして位置づけることができる。

研究成果の概要（英文）：A clustering algorithm is a fundamental tool for analyzing data set. Most algorithms are used distance between the feature vectors described from each data. However, a feature vector is always extracted. In this case, we describe a set as a feature. If we use a feature based on a set, we can not use clustering algorithms using distance. Thus, we need new clustering algorithm for using both similarity and distance.
A key idea of our clustering algorithm is to make mutually nearest neighbor graph (MNN-Graph). Our clustering algorithm consists of 5 steps; (1)Extract features from data set, (2)Make MNN-Graphwhich which regard data as vertexes, (3)Extract cliques in MNN-Graph, (4)Return to step (2) until a termination condition, (5)Combined similar sub-graph set,then output the result set as a cluster. We experimented on a set of document data.From experiments, we can say that accuracy of clustering was not so bad.

研究分野：データ工学

キーワード：クラスタリング　相互隣接グラフ　相互ランキング　類似度

1　研究開始当初の背景

ビックデータの活用を活用するため，データ解析技術，可視化技術，ビジュアルインタラクション技術の開発が必要である．特に，膨大な量のデータからデータ解析するためには，個々のデータを個別に分析することに加え，データを何らかの塊に分類し，分類した塊を分析することが必要である．すなわち，効果的かつ効率的にデータを取り扱うためには，クラスタリング技術が必要不可欠な技術である．現在，クラスタリング手法として，階層的クラスタリング，非階層的クラスタリング，自己組織化マップなどが用いられている．これらクラスタリング手法を実際のデータに適用する場合，データ間の類似性を測る尺度を決める必要がある．ここで，データ間の類似性を測る尺度として，１）距離の公理を満たすもの（以降，距離と呼ぶ），２）距離の公理を満たさないもの（以降，類似度と呼ぶ）があるが、その尺度の決定することにより，適用可能なクラスタリング手法が限定されてしまう．たとえば，類似性を測る測度として距離を用いた場合，階層的クラスタリング，非階層的クラスタリング，自己組織化マップのすべての手法が適用可能である．しかし，類似性を測る測度として類似度を用いようとした場合，類似度は順序尺度であることから，階層的クラスタリング，非階層的クラスタリング，自己組織化マップのすべてが適用不可能である．

2　研究の目的

非計量多次元尺度構成法は，非計量空間内のデータを，データ間の「類似性」の順序関係を保存するという条件の下で，ある計量空間（ユークリッド空間な ど）にデータを埋め込む手法である．このことは，非計量空間内のデータに対して非計量多次元尺度構成法を実行することにより，計量空間でのみ適用可能であった階層的クラスタリング，非階層的クラスタリング，自己組織化マップなどのクラスタリング手法が適用可能可能であることを示している．本研究の目的は，非計量多次元尺度法に依存することなく、問い合わせデータに対する検索順位に基づいたクラスタリング手法を提案することである。このことにより、データ間の類似性を測る尺度が距離・類似度のどちらを選択しても適用可能なクラスタリング手法となることを提案する。

3　研究の方法

3.1　提案手法の枠組み

クラスタリング手法の手順を図 1 に示す．はじめに，入力されたデータ集合の各データをグラフの頂点に対応させ，相互隣接グラフを生成する (図 1(1))．図 1(1) 中の円形の頂点はそれぞれ各データに対応した頂点であり，相互に類似している頂点間をリンクで接続している．これにより，頂点間のリンクの有無により DATA 間の類似性の関係を表現する．

次に相互隣接グラフから部分グラフを抽出する (図 1(2))．抽出する部分グラフとしてはクリークを用いる．クリークとはグラフ内の任意の頂点集合においてその全ての頂点間でリンクが接続されている部分的な完全グラフのことを指す．図 1(2) ではクリークを形成している頂点集合を太線および細線，太い破線，細い破線の四角い枠でそれぞれ囲んでいる．相互隣接グラフから抽出されたクリークの頂点集合に対応する文書集合は類似性の強い文書の集まりである．そのため抽出された各クリークはそれぞれクラスタの核となると考えられる．　そして抽出された各クリークをそれぞれ一つの頂点として扱い，再度相互隣接グラフを生成する (図 1(3))．ここで図 1(3) 中の四角の頂点は図 1(2) で抽出された各クリークをそれぞれ一つの頂点として扱ったものである．クリークを一つの頂点としたものと類似している頂点は，クラスタの核とした文書集合に類似している文書である．そして図 1(2) および図 1(3) を抽出される部分グラフ集合の変化が収束するまで繰り返し行い，類似する部分グラフを結合する．

最後に収束した部分グラフ集合からからクラスタを生成する (図 1(4))．部分グラフ抽出においてクリークを抽出した場合，一つの頂点が複数のクリークに属する可能性がある．また図 1(2) および図 1(3) を繰り返し行うため，ある頂点集合が複数の部分グラフに属する場合も存在する．そのため抽出された部分グラフ集合の中には類似した頂点集合を持つ部分グラフ同士が存在することがあるため，この類似している部分グラフ同士を結合するという処理を行う．この結合処理を行った結果をそれぞれ出力クラスタとする．図 1(3) では上部にクリーク，中部にクリークに属する 3 つの頂点のうち 2 つが同じである 2 つのクリーク，右下部に孤立点が存在している．図 1(4) では枠線で囲われたこれらの部分グラフを抽出し，それぞれ出力クラスタとする．

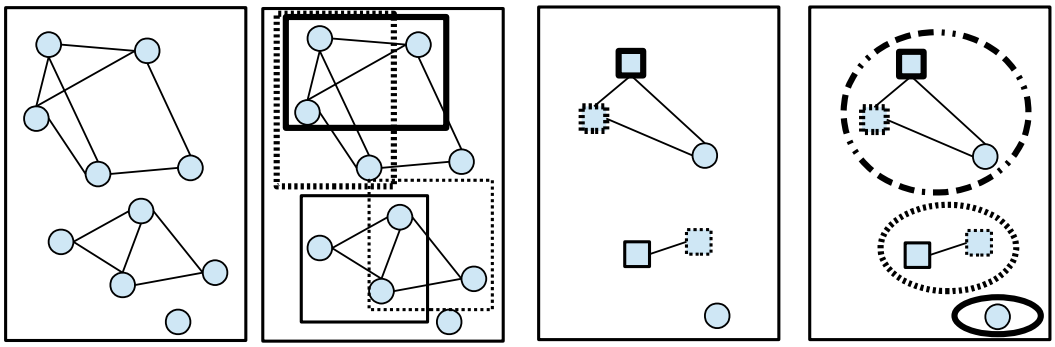

図 1: クラスタリング手法の手順

3.2 文書集合に対する相互隣接グラフ生成

M 件の文書集合 $\boldsymbol{D} = \{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_M\}$ が与えられ，各文書に出現する名詞の和集合 $\boldsymbol{\Lambda} = \{\lambda_1, \ldots, \lambda_L\}$ があるとする．このとき名詞 λ_l の出現頻度 $p(\lambda_l)$ として全文書中における λ_l の出現する文書数の割合を求める．

$$p(\lambda_l) = \frac{(\lambda_l\text{の出現する文書の数})}{L} \tag{1}$$

そして各文書 $\boldsymbol{D}_i$ において出現する名詞のうち $p_{\min} \leq p(\lambda_l) < p_{\max}$ を満たす名詞の集合 $\boldsymbol{n}_i$ を抽出し，文書集合 $\boldsymbol{D}$ に対応する名詞集合 $\boldsymbol{N} = \{\boldsymbol{n}_1, \ldots, \boldsymbol{n}_M\}$ を求める．ここで $p_{\min}$および $p_{\max}$ はそれぞれ出現頻度の下限と上限のしきい値である．

この名詞集合 $\boldsymbol{N} = \{\boldsymbol{n}_1, \ldots, \boldsymbol{n}_M\}$ およびグラフ $\boldsymbol{G} = \{\boldsymbol{V} \equiv \{v_1, \ldots, v_M\}, \underline{o}ldsymbol{E} \equiv \{\emptyset\}\}$ を入力とし，相互隣接グラフ $\boldsymbol{G_D}$ を出力としたときの相互隣接グラフ生成アルゴリズムを **MNN 1** ～ **5** に示す．

MNN 1: 文書割り当て

定義した $\boldsymbol{V}$ の各頂点 v_i に対して文書 $\boldsymbol{D}_i$ の名詞集合 $\boldsymbol{n}_i$ を割り当てる．

MNN 2: 類似度計算

頂点 v_i と v_j の類似度 $\mathrm{Sim}(v_i, v_j)$ を計算する．ここでは類似度として集合間の類似度を示す Jaccard 係数

$$\mathrm{Sim}(v_i, v_j) = \frac{|\ \boldsymbol{n}_i \cap \boldsymbol{n}_j\ |}{|\ \boldsymbol{n}_i \cup \boldsymbol{n}_j\ |} \tag{2}$$

を算出する．これを全頂点間で算出する．

MNN 3: 類似度順位算出

頂点 v_i に対する頂点 $v_j (j = 1, \ldots, M; j \neq i)$ の類似度の大きさを比較し類似度が大きいほど順位が高くなるように類似度順位 r_{ij} を計算する．ただし，r_{ij} は $j = i$ のとき $r_{ij} = 0$, $\mathrm{Sim}(v_i, v_j) = 0$ のとき $r_{ij} = \infty$ とする．これを全頂点間で求める．

MNN 4: リンク接続

r_{ij} と順位のしきい値 h をもとに $\boldsymbol{G}$ のリンク集合 $\boldsymbol{E}$ の (i, j) 要素 e_{ij} を次のように定義し，相互に順位が h 位以内の頂点同士でリンクを接続する．

$$e_{ij} = \begin{cases} 1 & r_{ij} \leq h \text{ かつ } r_{ji} \leq h \\ 0 & \text{それ以外} \end{cases} \tag{3}$$

MNN 5: 相互隣接グラフ出力

$\boldsymbol{G}$ を $\boldsymbol{D}$ に関する相互隣接グラフ $\boldsymbol{G_D}$ として出力する．

3.3 相互隣接グラフを用いた文書クラスタリング手法

文書集合 $\boldsymbol{D} = \{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_M\}$ に対する本稿のクラスタリングアルゴリズムを **CLT 1** ～ **10** に示す．

CLT 1: 名詞の出現頻度算出

M 件の文書集合 $\boldsymbol{D} = \{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_M\}$ に対し，各文書に出現する名詞の和集合 $\boldsymbol{\Lambda} = \{\lambda_1, \ldots, \lambda_L\}$ および各名詞 $\lambda_l \in \boldsymbol{\Lambda}$ の出現頻度 $p(\lambda_l)$ を求める．

CLT 2: 名詞集合抽出

各文書 $\boldsymbol{D}_i$ に出現する名詞のうちの $p_{\min} \leq p(\lambda_l) < p_{\max}$ を満たす名詞の集合 $\boldsymbol{n}_i$ を求め，文書集合 $\boldsymbol{D}$ に対する名詞集合 $\boldsymbol{N} = \{\boldsymbol{n}_1, \ldots, \boldsymbol{n}_M\}$ を求める．

CLT 3: 名詞集合に関する相互隣接グラフ生成

$\boldsymbol{N}$，初期グラフ $\boldsymbol{G}$，順位に関するしきい値 h を入力とし Procedure 1 に示した MNN$(\boldsymbol{N}, \boldsymbol{G}, h)$ を実行し，出力として相互隣接グラフ $\boldsymbol{G_D}$ を得る．

CLT 4: $\boldsymbol{G_D}$ から部分グラフ抽出

$\boldsymbol{G_D}$ からクリークおよび孤立点の集合 $\boldsymbol{C} = \{\boldsymbol{C}_1, \ldots, \boldsymbol{C}_K\}$ を得る．ここで $\boldsymbol{C}$ の要素 $\boldsymbol{C}_a (a = 1, \ldots, K)$ は部分グラフを構成する頂点集合もしくは孤立点である．

CLT 5: 部分グラフの名詞併合

$\boldsymbol{C} = \{\boldsymbol{C}_1, \ldots, \boldsymbol{C}_K\}$ の要素 $\boldsymbol{C}_a (a = 1, \ldots, K)$ に対して新たに $\boldsymbol{C}_a$ の名詞の集合 $\boldsymbol{n}_{\boldsymbol{C}_a}$ を求め，$\boldsymbol{C}$ の名詞集合 $\boldsymbol{N_C} = \{\boldsymbol{n}_{\boldsymbol{C}_1}, \ldots, \boldsymbol{n}_{\boldsymbol{C}_K}\}$ とする．

CLT 6: 部分グラフに関する相互隣接グラフ生成

$\boldsymbol{N_C}$，初期グラフ $\boldsymbol{G}$，順位に関するしきい値 h を入力とし Procedure 1 に示した MNN$(\boldsymbol{N_C}, \boldsymbol{G}, h)$ を実行し，出力として相互隣接グラフ $\boldsymbol{G_C}$ を得る．

CLT 7: $\boldsymbol{G_C}$ から部分グラフ抽出

$\boldsymbol{G_C}$ から新たにクリークおよび孤立点の集合 $\boldsymbol{C} = \{\boldsymbol{C}_1, \ldots, \boldsymbol{C}_K\}$ を得る．

CLT 8: 繰り返し処理

抽出される部分グラフが収束するまで **CLT 5**〜**CLT 7** を繰り返す．収束した結果の $\boldsymbol{C}$ を $\boldsymbol{C}^\star = \{\boldsymbol{C}_1^\star, \ldots, \boldsymbol{C}_{K^\star}^\star\}$ とする．

CLT 9: 部分グラフのマージ処理

$\boldsymbol{C}^\star$ 内の任意の 2 つの要素 $\boldsymbol{C}_a^\star, \boldsymbol{C}_b^\star (a \neq b)$ に対して類似度

$$\mathrm{Sim}_{\boldsymbol{C}}(\boldsymbol{C}_a^\star, \boldsymbol{C}_b^\star) = \frac{|\ \boldsymbol{C}_a^\star \cap \boldsymbol{C}_b^\star\ |}{\min(|\ \boldsymbol{C}_a^\star\ |, |\ \boldsymbol{C}_b^\star\ |)} \tag{4}$$

を計算する．この類似度は $\boldsymbol{C}_a^\star$ と $\boldsymbol{C}_b^\star$ の間の頂点の重なりの度合いを示している．完全に一致することはないため $0 \leq \mathrm{Sim}(\boldsymbol{C}_a^\star, \boldsymbol{C}_b^\star) < 1$ である． $\mathrm{Sim}(\boldsymbol{C}_a^\star, \boldsymbol{C}_b^\star) \geq \epsilon$ のとき，$\boldsymbol{C}_a^\star$ と $\boldsymbol{C}_b^\star$ を結合する．この処理を $\boldsymbol{C}^\star$ が収束するまで結合を繰り返す．

CLT 10: クラスタ出力

収束した $\boldsymbol{C}^\star$ をクラスタ集合 $\boldsymbol{\mathcal{C}}$ として出力する．

4 研究成果

4.1 文書クラスタリング

文書数 100，クラス数 10 となるデータ集合を生成し、クラスタリング手法の評価を実施した．まずはじめに，3.2 節の相互隣接グラフの生成方法に基づき文書集合から相互隣接グラフを生成した．図 2 は、文書集合に対する初期の相互隣接グラフを表す．初期の相互隣接グラフ（図 2）では，クリークの数が 10, 孤立点の数が 15 であることが分かる．順位のしきい値 h によりクラス生成の影響について検討した．順位のしきい値 $h = 5, 6$ で生成されたクラスタ集合において各クラスタに属する文書のリストを表 1 に示す．表 1 の各クラスタに属する文書の文書番号を示しており，太字で強調されている文書番号は同一文書とは異なるの文書を示している．$h = 5, 6$ での孤立点の数は，それぞれ 3 と 6 であり，誤分類の総数は，それぞれ 5 と 1 であったことが分かる．これらのことから分類に失敗した数（誤分類と孤立点の合計）は、それぞれ $8 (= 3 + 5)$ と $7 (= 6 + 1)$ であったと言える．すなわち，クラスタリングの成功率は、それぞれ 0.92 と 0.93 である．

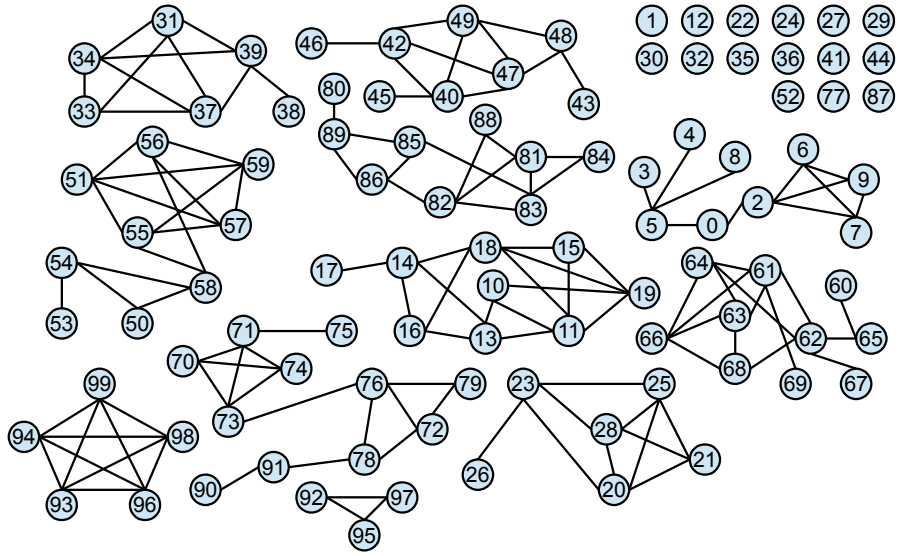

図 2: 初期相互隣接グラフ

表 1: $h = 5, 6$ におけるクラスタに属する文書のリスト

クラスタ番号 k	しきい値	
	$h = 5$	$h = 6$
1	0, 1, 2, 3, 4, 5, 6, 7, 8, 9; 計 10	0, 1, 2, 3, 4, 5, 6, 7, 8, 9; 計 10
2	10, 11, 12, 13, 14, 15, 16, 17, 18, 19; 計 10	10, 11, 12, 13, 14, 15, 16, 17, 18, 19; 計 10
3	20, 21, 22, 23, 25, 26, 27, 28; 計 8	20, 21, 23, 24, 25, 26, 28, 29; 計 8
4	30, 31, 32, 33, 34, 35, 36, 37, 38, 39; 計 10	30, 31, 32, 33, 34, 35, 37, 38, 39; 計 9
5	40, 41, 42, 43, 44, 45, 46, 47, 48, 49; 計 10	40, 41, 42, 43, 44, 45, 46, 47, 48, 49; 計 10
6	50, 51, 52, 53, 54, 55, 56, 57, 58, 59, **60**; 計 11	50, 51, 52, 53, 54, 55, 56, 57, 58, 59, **60**; 計 11
7	61, 62, 63, 64, 65, 66, 67, 68, 69; 計 9	61, 62, 63, 64, 66, 67, 68, 69; 計 8
8	**62**, **64**, **67**, **68**, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79; 計 14	70, 71, 72, 73, 74, 75, 76, 78, 79; 計 9
9	80, 81, 82, 83, 84, 85, 86, 87, 88, 89; 計 10	80, 81, 82, 83, 84, 85, 86, 88, 89; 計 9
10	91, 92, 93, 94, 95, 96, 97, 98, 99; 計 9	90, 91, 92, 93, 94, 95, 96, 97, 98, 99; 計 10
孤立点	24, 29, 90; 計 3	22, 27, 36, 65, 77, 87; 計 6

4.2 マルウェア検体に対するクラスタリング

研究期間中にマルウェアの動的解析結果から特徴量を抽出し，提案アルゴリズムの適用を試みた．しかし、入手したマルウェアはほとんど同一種であり実験には至らなかった．今後マルウェアの収集を行い本提案手法を適用予定である．

参考文献

[1] 田中善弘，大野克嗣，横山和成，”非計量多次元尺度法への期待と新しい視点”，総計数理，Vol. 49, No.1, pp. 133-153, 2001.

[2] 熊谷敦也，”古典的他事点尺度構成法の視点からの関連性データの非対称性への対処”，日本応用数理学会論文誌，Vol. 21. No. 2, pp. 165-174, 2011

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考