行列上の相互順序決定

上原子 正利[†] 小柳 滋[†]

本論文で我々は，行列上の 1 つの非ゼロ要素に対する類似度に基づいて行の順序集合と列の順序集合のペアを互いに依存するように決定するアルゴリズムを定義する．行あるいは列の一方の順序集合を決定するアルゴリズムには文書処理におけるベクトル空間法があり，本論文のアルゴリズムはこれを特殊例として含む．このアルゴリズムは 2 値疎行列から密部分行列を発見する 2 値内積縮退 MC から派生している．我々はアルゴリズムの定義に加え，このアルゴリズムを実際の文書データに適用することでその性質の確認を行う．

Co-ordering on a Matrix

Masatoshi Kamiharako[†] and Shigeru Oyanagi[†]

We define an algorithm which determines on a matrix a pair of rows and columns ordered sets in a mutually dependent manner. Those orders are decided based on similarity to a non-zero element on the matrix. The algorithm includes as a special case vector-space model of text processing which determines an ordered set of either rows or columns. The algorithm is derived from binary inner-product degeneration MC which finds a dense submatrix from a sparse binary matrix. Besides defining the algorithm, we confirm its characteristics by use of a real text dataset.

1. はじめに

データマイニングの問題で用いられるデータセットの多くは行列として表現でき，それらに対する操作は一般に行に対するものと列に対するものを組み合わせている．そのような問題の例に文書処理における関連文書決定がある．このときの処理は，文書を行，単語を列，各文書での各単語の出現回数を要素の値として行列を構成したとき，まず各列ベクトルの要素の値をその非ゼロ要素数に基づいて調整する．次に行ベクトル間の内積を求め，その値を 2 つのベクトルのノルムの積で割るなどして調整したものを類似性の尺度として関連文書を決定する．

多くの場合，このようなアルゴリズムは行か列の一方のみを出力する．たとえば関連文書決定の出力は行だけである．それに対して，行と列を同時に出力するアルゴリズムもいくつか存在する．たとえばピンポンアルゴリズム[1] は，2 値疎行列から一定の条件を満たす密部分行列を 1 つ取り出す．また，二部スペクトルグラフ分割アルゴリズム[2] は行列の相互クラスタリング，すなわち行と列が相互に依存するようなクラスタリングを行うことで，行と列の部分集合のペアを複数出力する．

本論文で我々は，これらと同じく行と列を同時に出力するアルゴリズムを定義する．このアルゴリズムは，行列中の 1 つの非ゼロ要素に対する類似度に基づいて，行の順序集合と列の順序集合のペアを互いに依存するように決定する，すなわち，このアルゴリズムは行と列の相互順序決定を行う．このアルゴリズムは 2 値疎行列から密部分行列を発見する 2 値内積縮退 MC[3] から派生している[☆]．

本論文の構成は次のとおりである．まず 2 章で，行と列の操作を組み合わせる各種のアルゴリズムを整理し，その中に本論文で最終的に定義するアルゴリズムを位置付ける．3 章では 2 値内積縮退 MC を概観し，4 章でこれを一般化した縮退 MC を定義する．5 章でその処理過程を利用して相互順序決定を行う縮退 MC 削除部分再構成を定義し，6 章で実際のデータへの適用例を示す．最後に 7 章で問題点を議論する．

† 立命館大学
Ritsumeikan University

☆ 文献 3) でのこのアルゴリズムの名称には「2 値」が付いていない．この変更は本論文での拡張を反映したものである．

2. 行と列の操作を組み合わせるアルゴリズム

行と列の操作を組み合わせるアルゴリズムは，出力が行と列の一方か両方かで 2 つに分けられる．本章で我々はこれらを概観し，本論文で定義するアルゴリズムの位置付けを行う．

2.1 行か列の一方のみを出力するアルゴリズム

行と列の操作を組み合わせて行か列の一方のみを出力するアルゴリズムの例に，前述の関連文書決定方法がある．このアルゴリズムは一般にベクトル空間法と呼ばれる[4]．これが行と列の操作を組み合わせる理由は，対象となる文書行列の性質にある．その性質とは，意味的に重要でない要素が大きな値を持ちうるというものである．このような要素として，頻出語に相当する列ベクトルや長文に相当する行ベクトルのものがある．そのため，このような要素が類似性の高低を左右しないように行と列の両方を調整する必要がある．

ベクトル空間法での列ベクトルの調整は，IDF と呼ばれる値を乗じることで行われる．以下ではこの操作を IDF 調整と呼ぶ．IDF の定義はいくつかあるが，たとえばその列ベクトル中の非ゼロ要素数を df，行列の総行数を N とすると，$\log(N/df)+1$ である[4]．行ベクトル間の類似性尺度の例として，内積をノルムの積で割って調整したものがある．以下ではこの調整をノルム 1 と呼ぶ．内積とノルム 1 の組合せは余弦になる．

この種のアルゴリズムの他の例に，2 値行列を扱う相関規則決定アルゴリズム[5),6)] がある．ユーザを行，アイテムを列，要素をユーザのアイテム購買履歴とした行列を用いると，相関規則とはある条件を満たす 2 つのアイテム集合の順序付きペアである☆．その条件とは，自然数 s と 1 以下の正の実数 c を与えられたとき，2 つのアイテム集合に相当する列部分集合の間の条件付き確率が c 以上で☆☆，2 つの列部分集合すべてに 1 要素を持つ行の数が s 以上というものである．行と列の条件を併用するため，ここでも行と列の操作の両方が必要になる．

さらに，推薦システム[8)] の GroupLens[9)] もこの種のアルゴリズムを用いる．ここでの行列は，行を文書，列をユーザ，要素をユーザの文書に対する評価値としたものである．このアルゴリズムは 1 つの列を指定されると，その列の空欄の値を推定し，その値の高い行を出力する．値を推定するために，列ベクトル間の相関係数を計算し，それを用いて空欄の行の非ゼロ要素を重み付けした和を求める．ここでも行と列の操作が併用されている．

2.2 行と列の両方を出力するアルゴリズム

行と列の両方を出力するアルゴリズムのうち単純なものは，一方のみを出力するアルゴリズムにわずかな変更を加えて実現できる．たとえば Columbia Newsblaster の文書整理アルゴリズム[10)] は，ベクトル空間法で行部分集合を決定し，その集合中に非ゼロ要素を持つ列から品詞情報などを用いて特徴的なものを取り出す．列部分集合も取り出す目的は，文書群を説明する単語群の決定にある．

Newsblaster のアルゴリズムは行と列の決定を別々に行うが，他のいくつかのアルゴリズムは行と列の決定を処理過程で相互に依存させる．たとえば，2 値疎行列から密部分行列を発見するピンポンアルゴリズム[1)] は，行部分集合からの列部分集合の選択と列部分集合からの行部分集合の選択を繰り返し，行と列が相互に選択しあう部分行列を出力する．この部分行列は関連する行と列のペアを表す．

ピンポンは行部分集合と列部分集合のペアを 1 つだけ出力するが，複数を出力するアルゴリズムもある．それらは行と列を相互に依存するようにクラスタリングするため，その操作は相互クラスタリングと呼ばれる．相互クラスタリングを行う二部スペクトルグラフ分割[2)] は，ベクトル空間法での IDF 調整を行った行列から文書の部分集合とそれを説明する単語の部分集合のペアを複数決定する．

他の相互クラスタリングアルゴリズムに RPSA[11)] がある．これはピンポンと同様に疎行列から密部分行列を発見する操作を基本としているが，ピンポンと異なり対象が多値行列である．そして，発見した密部分行列を段階的に併合することで階層的クラスタリングを行う．

2.3 本論文のアルゴリズムの位置付け

我々が本論文で最終的に定義するアルゴリズムは，行と列に同等の操作を適用し，行と列の両方を出力する．この点ではピンポンなどと同様であるが，それらが部分集合ペアを決定するのに対して，このアルゴリズムは互いに依存する行の順序集合と列の順序集合のペアの決定，すなわち行と列の相互順序決定を行う．これまで述べたものではベクトル空間法だけが順序集合を出力するが，本論文のアルゴリズムはこれを特殊例として含む．

応用の観点からの本論文のアルゴリズムの目的は 2

☆ 文献 5) の定義では 1 つのアイテム集合と 1 つのアイテムだが，文献 6) で 2 つのアイテム集合に拡張されている．

☆☆ この値は信頼値と呼ばれるが，これが条件付き確率であることは文献 7) で指摘されている．

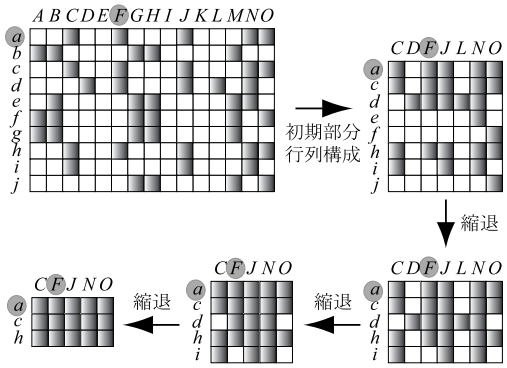


図 1 BIpDMC
Fig. 1 BIpDMC.

つある．1 つは，相互クラスタリングなどと同様に，意味的に関連する行とそれらを説明する列を同時に得ることである．もう 1 つは，ベクトル空間法のような行ベクトル全体からの類似行の発見ではなく，特定の列に注目したときの類似行の発見である．行と列を入れ替えても同様の議論が成立する．

このアルゴリズムは 2 値内積縮退 MC[3] から派生したものである．以下ではそのいくつかの制約を取り除いて一般化し，さらにその処理過程を利用して相互順序決定を行うアルゴリズムを定義する．

3. 2 値内積縮退 MC

2 値内積縮退 MC（以下 BIpDMC とする）は，2 値疎行列とその中の 1 つの 1 要素，2 つの基本操作の関数，および終了条件を指定されることで密部分行列を出力するアルゴリズムである．開始時に指定される 1 要素を基準要素，その行を基準行，列を基準列と呼ぶ．2 つの基本操作の 1 つは初期部分行列構成であり，これは基準要素を元に疎でサイズの大きい部分行列を構成する．もう 1 つは縮退操作であり，これは基準行・列との内積の小さい行・列を，指定された終了条件を満たすまで段階的に部分行列から削除してゆく．終了条件には部分行列の密度，行数，列数，面積などがある．**図 1** に (a, F) を基準要素としたときの実行例を示す．以下ではこれらの操作を概観し，さらにこのアルゴリズムの処理過程を 2 通りの方法で表現する．

3.1 初期部分行列構成

初期部分行列構成は基準要素と弱い関係を持つ行と列からなる疎でサイズの大きい部分行列を構成する操作である．この構成方法はいくつかあるが，最も単純なものは 2 つある．1 つは基準行・列およびそれらと共通要素を持つすべての行・列を初期行・列集合とするものである．このときの集合をそれぞれ R_l，C_l と呼ぶ．もう 1 つの方法は，基準列・行に 1 要素を持つすべての行・列を初期行・列集合とするものである．このときの集合をそれぞれ R_s，C_s と呼ぶ．これらの組合せで 4 種類の初期部分行列が構成されうる．

R_l を実際に用いる場合は次の問題を考慮する必要がある，その問題とは，R_l が実質的な関連のない行を初期行とする場合があり，それらの行が後述の縮退操作の計算コストを浪費することである．例として WWW サイトのアクセスログから構成される行列を考える．行をセッション，列をページ，要素をアクセスの有無とすると，多くの行がサイトのトップページに相当する列に 1 要素を持つため，R_l はほとんどの行を含む．このような場合に不要な行を排除するには，1 要素数が一定以下の列に共通要素を持つ行のみを，あるいは R_l のうち基準行との共通要素の多い行のみを初期行とする必要がある．C_l についても同様である．以下ではこのような操作を初期部分行列のサイズ制御と呼ぶ．

3.2 縮退操作

縮退操作はスコア計算と削除操作からなり，それぞれ行方向と列方向の 2 つがある．行方向のスコア計算は，部分行列中の各行について，基準行との類似性を示すスコアである内積を求める操作である．対象が 2 値行列であるため，2 つの行の内積はそれらの共通要素数に等しい．行方向の削除操作は，スコアの小さい行を部分行列から削除する操作である．このとき，スコア最小の行だけを削除するなら，その縮退操作を最小値縮退と呼ぶ．さらに，最小スコア行以外も同時に削除すると縮退操作の回数が減り，処理が高速化される．同時に削除する行数の目安として，その時点での行集合のサイズに対する割合を用いることができる．この割合を縮退率と呼び，このときの縮退操作を割合縮退と呼ぶ．なお，同じスコアのものはすべて残すかすべて削除する．列方向についても同様である．

縮退操作の際には，行と列のどちらを扱うか，あるいはどちらも縮退せずに終了するかを決定する必要がある．これを行う関数を縮退方向決定関数と呼ぶ．この関数の定義によって，基準要素などの他の入力が同じでも出力部分行列が異なってくる．BIpDMC はこの関数の具体的な構成方法を指定しない．なぜなら，この関数の構成方法によって処理速度と出力部分行列に対する要求を満たす程度が変わるが，対象問題によって許容できる処理時間や出力に対する要求の厳密さが異なるためである．

3.3 処理過程の 2 つの表現

BIpDMC の処理過程はグラフとして表現できる．こ

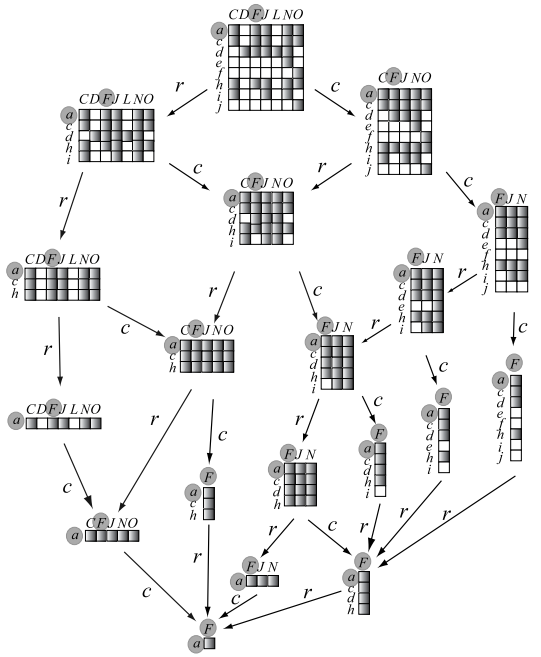


図 2 縮退グラフ
Fig. 2 A degeneration graph.

のグラフは，部分行列を節点，最小値縮退による部分行列の変化を 1 つの枝とした有向グラフであり，縮退方向決定関数の出力の可能なすべての組合せとその結果得られるすべての部分行列を含む．このグラフを縮退グラフと呼び，グラフ上の経路を縮退経路と呼ぶ．**図 2** は図 1 を含む縮退グラフの例である．基準要素は (a, F) であり，r と c はそれぞれ行縮退と列縮退を示す．この表現を用いると，BIpDMC の処理過程は縮退グラフ上で望ましい節点を探索するグラフ探索と見なせる．最小値縮退は枝を 1 つずつ進むことに相当し，割合縮退は r または c の枝を一度に 1 つ以上，初期部分行列に近いほど大きく進むことに相当する．

BIpDMC の処理過程は基準行に対する類似行決定の観点からも表現できる．そのためにまず列縮退を行わない類似行決定と列縮退を限界まで行った後の類似行決定を考え，その後それらの中間を考える．列縮退を行わないときの類似行決定は，初期部分行列全体から類似行を決定することに等しい．列縮退だけを限界まで行った後の類似行決定は，基準列に 1 要素を持つ行を選ぶことに等しい．これらの中間はある程度の列縮退を行った後の類似行決定であり，これは基準列とある程度類似した列にのみ注目して類似行決定を行うことに相当する．このため，BIpDMC の処理過程は，類似行決定のために注目する列を段階的に限定してゆく過程と表現できる．このとき，基準列が異なれば処理過程で注目する列も異なり，得られる類似行も異なる．同様の議論が類似列決定についても成立する．

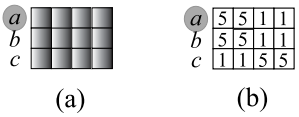


図 3 2 値と多値の違い
Fig. 3 Difference between binary and multi-valued.

4. 縮 退 MC

BIpDMC は 2 つの点で一般化できる．1 つは処理対象を 2 値行列から一般の多値行列に拡張することであり，もう 1 つは類似性の尺度に内積以外を認めることである．BIpDMC を一般化して，2 値および多値行列上で類似性尺度を任意としたアルゴリズムを縮退 MC（以下 DMC）と呼ぶ．BIpDMC は DMC の一種である．

4.1 多値行列への拡張

2 値行列から多値行列への拡張が必要になるのは行列が密な場合である．疎であれば非ゼロ要素の存在だけで関連性を表現できるが，密であれば存在だけでは不十分で，値が重要になる．たとえば，**図 3** (a) の 2 値行列では a に対する b と c の類似度が等しくなるが，この行列が実際には図 3 (b) のように多値なら b と c の類似度は異なる．

4.2 類似性尺度の一般化

類似性尺度は行列を 2 値から多値にすることで変わるが，2 値行列に対しても内積以外のものを利用できる．以下ではまず 2 値行列に対する尺度を述べ，その後に多値行列に対する尺度を述べる．さらに，尺度の一般化によって初期部分行列の種類が増えるため，その構成方法も述べる．

4.2.1 2 値行列での類似性尺度

行列中の 1 要素数の分布によっては，共通要素数を類似性尺度として用いることが不適切になる．たとえば前述の WWW アクセスログの行列では，列ベクトルの 1 要素数にばらつきがあり，1 要素数の多い列は少ない列より基準列に類似していると判定される傾向を持つ．

この問題を解決するには，共通要素だけでなく非共通要素の数も類似性の尺度に反映させればよい．そのような方法として，ベクトル空間法と同様のノルム 1 が利用できる[12]．他にも，ベクトルを 1 要素の集合と見なしたときの和集合の要素数や，片方のベクトルの全 1 要素数による調整が考えられる．これらの関係を**図 4** の A と B を用いて考えると次のよう

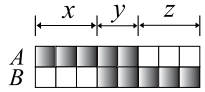


図 4 2 つの 2 値ベクトル
Fig. 4 Two binary vectors.

になる．内積での類似度は y となり，ノルム 1 では $y/(\sqrt{x+y}\sqrt{y+z})$ になる．和集合の要素数による調整は $y/(x+y+z)$ となり，これは 2 値ベクトルの Jaccard 係数[4] である．片方のベクトルの全 1 要素数による調整は $y/(x+y)$ または $y/(y+z)$ となり，これらはそれぞれ条件付き確率 $P(B|A)$ と $P(A|B)$ である．

4.2.2 多値行列での類似性尺度

多値行列での類似性尺度も 2 値行列と同様にいくつか考えられ[4]，たとえばベクトル空間法での内積や GroupLens と同様の相関係数などがある☆．ただし，多値行列に対しては次の処理が重要になる．それは，行縮退のときの列の調整，および列縮退のときの行の調整である．以下では，行縮退のときの行を主方向，列を副方向と呼び，同様に列縮退では列を主方向，行を副方向と呼ぶ．このとき，この操作は副方向調整と表現でき，前述の 2 値ベクトルのノルム 1 などは主方向調整と表現できる☆☆．

副方向調整はベクトル空間法における列ベクトルの IDF 調整を一般化したものであるが，すべての行列で副方向の IDF 調整が望ましいとは限らない．たとえば，密な行列を扱う場合には IDF 調整が有効に働かないため，ノルム 1 のような要素の値に注目した調整を行う必要がある．

なお，調整操作の際には行列の要素の値を変更せず，スコア計算時の一時変数だけを変更する．要素の値を変更すれば，削除された行や列の影響が削除されなかった要素に残り，同じ行と列からなる部分行列の要素の値が縮退経路によって異なる．このような煩雑さを避けるため，要素の値は変更しない．

4.2.3 初期部分行列の新しい構成方法

類似性尺度の一般化によって初期部分行列の新しい構成方法も導入される．その方法を**図 5** で説明する．

基準要素 (a, A) から $R_l \times C_l$ を構成した場合，$\{a,b,c,d\} \times \{A,B\}$ が得られるが，ここで c と d が問題になる．これらは元の行列で非共通要素を考慮すると a に対して異なるスコアを持つが，C_l の中では同じスコアを持つ．これらを初期部分行列内でも区別するには，列集合を R_l の各行の非ゼロ要素すべての列からなる集合とする必要がある．この列集合を C_{sl} と呼ぶ．これは図 5 で $\{A,B,C,D\}$ となる．行集合についても同様に R_{sl} を定義できる．

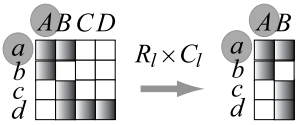


図 5 $R_l \times C_l$ では不十分な場合
Fig. 5 The case that $R_l \times C_l$ is insufficient.

R_l，R_s，R_{sl} の関係は，行列を行集合と列集合からなる二部グラフの隣接行列としてとらえると次のようになる．R_l は基準行および基準行から距離 2 のすべての節点（二部グラフなのですべての行），R_s は基準行および基準行から基準列を経由する距離 2 のすべての節点，R_{sl} は基準行および基準行から基準列を経由する距離 2 と 4 のすべての節点に相当する．C_l，C_s，C_{sl} についても同様である．

類似性尺度によって初期部分行列の構成方法が変わることは，両者に関係があることを示している．たとえば，行の内積は基準行から距離 1 の列しか考慮しないため，基準行から距離 3 の列，すなわち $C_{sl} - C_s$ は不要である．

図 5 はまた，初期部分行列のサイズ制御方法に 3.1 節以外のものが必要であることを示している．3.1 節では共通要素数だけを考慮したが，その方法では b, c, d を区別できない．d だけを排除するには，共通要素数だけでなく非共通要素数も考慮する．なお，要素の数だけでなくそれらの値まで見てスコアを計算することも考えられるが，その方法は処理時間を消費する．サイズ制御の目的は処理の高速化であるため，厳密で処理時間を要する計算より，大まかでも高速な計算が望ましい．

4.3 アルゴリズム

DMC のアルゴリズムは上記の点を除いて BIpDMC と変わらず，次のようになる．

(1) 行列中の非ゼロ要素を 1 つ指定し，その行を r_b，その列を c_b とする．
(2) r_b と c_b を用いて初期部分行列 M_0 を作る．M_0 の行集合を R_0，列集合を C_0 とする．
(3) $i = 0, 1, 2, \ldots$ について以下を繰り返す．
  (a) 縮退方向決定関数が停止を出力すれば，M_i を返して終了する．それ以外なら以下を行う．こ

☆ 文献 9) での GroupLens の方法は 2 つのベクトルの共通要素にのみ注目するため，非共通要素を考慮しない．そのため，この方法が有効になるのは行列が密な場合である．

☆☆ 2 値行列に対しても副方向調整を適用できる．たとえば文献 12) の BIpDMC では，副方向の基準ベクトルの 1 要素を 0 と見なす調整をしている．しかし，要素の値が大きく異なりうる多値と比べて，2 値では副方向調整の必要性が低い．

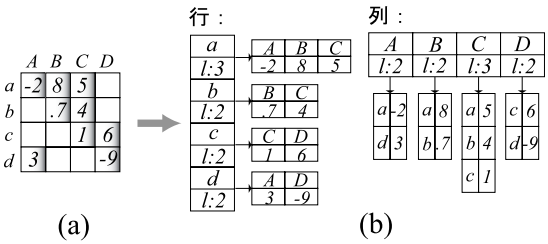


図 6 行列のデータ構造
Fig. 6 Data structure for a matrix.

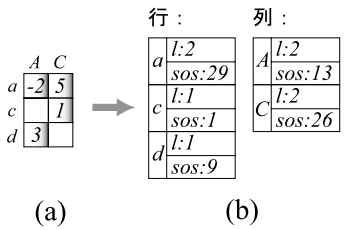


図 7 部分行列のデータ構造
Fig. 7 Data structure for a submatrix.

こでは行の場合を示すが，列でも同様である．

(b) M_i の各行のスコアを決める．

(c) R_i からスコアの小さい行をいくつか削除したものを R_{i+1} とする．その際，同じスコアのものはすべて削除するかすべて残す．

(d) C_i を C_{i+1} とする．

(e) $R_{i+1} \times C_{i+1}$ を M_{i+1} とする．

BIpDMC と同じく DMC も縮退方向決定関数の実装を指定しないため，ここではこの関数と縮退操作が分けられている．しかし，実装上で処理速度を考慮すると，この 2 つが 1 つにまとめられる[3]．

4.4 データ構造

DMC の重要なデータ構造は 2 つある．1 つは行列全体，もう 1 つは処理中の部分行列のものである．これらはいずれも文献 3) のデータ構造に改良を加えて得られる．

4.4.1 行　　列

図 6 (a)，(b) はある行列とそのデータ構造である．このデータ構造は，行列を行ベクトルの集合としてとらえたものと列ベクトルの集合としてとらえたもののペアである．文献 3) は各ベクトルを表現するために 1 要素の位置を要素とするソート済み配列を用いているが，ここでは多値データへの対応と処理の高速化のため，ハッシュテーブルによる連想配列を用いる．配列のインデックスは非ゼロ要素の位置，配列の値は非ゼロ要素の値である．また，各ベクトル中の非ゼロ要素数も記録し，初期部分行列のサイズ制御で利用する．図中の l がその値を示す．

4.4.2 部 分 行 列

図 7 (a)，(b) はある部分行列とそのデータ構造である．このデータ構造は行番号の集合と列番号の集合のペアである．文献 3) ではこれらの集合の表現にソート済み配列を用いたが，ここでは高速化のためハッシュテーブルを用いる．さらに，調整操作の高速化のため，各行・列ごとに部分行列内での要素の個数，要素の値の二乗和を記録しておく．この値は初期部分行列構成時に初期値が決定され，削除操作のたびに更新される．図中の l はベクトル中の非ゼロ要素数を，sos は値の二乗和をそれぞれ示す．

5. 縮退 MC による相互順序決定

DMC には適切な終了条件の設定が難しいという実用上の難点がある．たとえば，部分行列の密度は行集合と列集合の関連の強さを示すが，この値が高ければ精度が高まる一方で，再現率，すなわち関連するものを網羅する程度が低くなり，密度が低ければ逆になる可能性がある．この難点は DMC が初期行と初期列の部分集合を出力することからくる．

これを解決するため，我々は DMC の処理過程を利用して新しいアルゴリズムを定義する．このアルゴリズムは，非順序集合である初期行・列集合を基準要素に対する関連の強さによって決定される順序集合に変換する．この変換は，DMC の早い段階に削除される行・列ほど基準要素と関連が弱いと見なせることを利用し，行と列をそれぞれ削除された順の逆に並べることで実現される．その際，同時に削除された複数の行・列の順序の決定には削除時のスコアを用い，出力部分行列の行と列は順序集合の先頭に置く．このように決定される行と列の順序集合のペアを縮退 MC 順序と呼び，このアルゴリズムを縮退 MC 削除部分再構成（DMC Deletion Reconstruction，以下 DMCDR）と呼ぶ☆．DMCDR は相互順序決定を行う．

DMCDR の主要なデータ構造は次のとおりである．

`Del` 2 つの要素からなる構造体．1 つは縮退方向を示す 2 値変数 `roc`．もう 1 つは削除された番号と削除時のスコアのペアの配列 `array`．配列はスコアで降順ソートされる．

`stk` `Del` のインスタンスを積むためのスタック．開始時は空．

`ro, co` 行と列それぞれの番号の順序集合を表現するリスト．開始時は空．各番号には順位を示す数値

☆ 2 値行列を対象にした縮退 MC 順序は文献 12) ですでに述べられているが，そのアルゴリズムは本論文といくつかの点で異なる．そのうち 1 つは重要であるため，最終章で述べる．

を付随させる．

`osm` DMC の出力部分行列．

これらのデータ構造を用いた DMCDR のアルゴリズムは次のとおりである．

(1) DMC を実行して `osm` を得る．その際，削除のたびに `Del` のインスタンスを `stk` に積んでゆく．

(2) `osm` について，最後の縮退方向以外のスコアを計算する．

(3) `osm` の行番号をスコアの大きいものから順位とともに `ro` の終端に追加する．基準行のスコアはつねに最大とする．列についても同様の処理を行う．

(4) `stk` が空になるまで以下を繰り返す．

(a) `stk` のトップ `t` の `Del` の `roc` が行なら，`t` の `array` の番号を先頭から順位とともに `ro` の終端に追加してゆく．`roc` が列なら，同様に `co` の終端に追加してゆく．

(b) `t` を `stk` からポップする．

(5) `ro` と `co` を返して終了する．

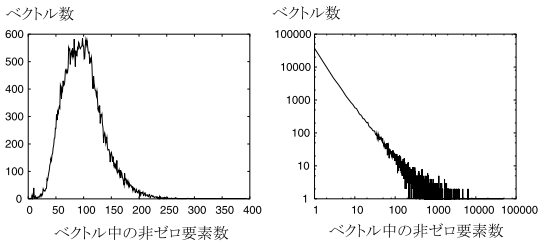
DMCDR は特殊例としてベクトル空間法を含む．DMCDR をベクトル空間法として実行する場合，副方向調整は IDF，基準要素は基準行内の任意の非ゼロ要素，初期部分行列は $R_{sl} \times C_{sl}$，縮退方向決定関数はつねに行縮退を出力，終了条件は最小行数 1 とする．あるいは，終了条件を最小密度 0 として一度もスコア計算をせずに DMC を終了すると，(2) で行のスコアが計算される．これらの方法で得られる `ro` がベクトル空間法の出力に相当する☆．

以上で本論文のアルゴリズムの定義は終了した．次にこのアルゴリズムを実際のデータに適用し，得られる出力を確認する．

6. 文書データに対する適用例

本章では DMCDR を実際のデータに適用して出力を確認する．用いるデータには出力の意味を把握しやすい文書を選ぶ．データセットは UCI KDD Archive に提供されている NSF Research Awards Abstracts ☆☆ の 1990 年から 94 年の部分である．

このデータから以下のように行列を構成する．まず，各文書からタイトル部と概要部の文字列を抜き出す．概要部が空の文書は無視する．次に，それらの文字列を空白で単語に分けて小文字に変換し，語幹を取り出す．最後に，各語幹の 1 つの文書内での出現回数を記録する．これによって，1 つの文書に相当する行と，その中の語幹に相当する非ゼロ要素列が得られる．語幹処理には Python のライブラリである nltk.stemmer.porter ☆☆☆ を用いた．なお，文書処理で一般に行われる冠詞などの不要語の除去は行わない．これは，領域知識によるデータの選択ができない場合を想定したためである．



(a) (b)

図 8 非ゼロ要素数の分布
Fig. 8 Distributions of non-zero elements.

この処理で得られた行列は，行数 49,078，列数 71,811，非ゼロ要素数 4,876,169，密度 1.4×10^{-3}，非ゼロ要素数の 1 行あたりの平均 99，1 列あたり 68 である．この行列の非ゼロ要素数の分布を**図 8** に示す．(a) と (b) は行列をそれぞれ行ベクトルの集合と列ベクトルの集合と見たときのものである．横軸はベクトル中の非ゼロ要素数，縦軸はベクトル数であり，(b) は両対数グラフである．

6.1 同一基準行内での基準列の変化の影響

まず，1 つの基準行の中で基準列が変化することによる出力の変化を見る．基準行は行列中の最初の行とする．これに相当する文書は a9000006『CRB: Genetic Diversity of Endangered Populations of Mysticete Whales: Mitochondrial DNA and Historical Demography』であり，**図 9** の語幹を含む．このうち，文書の意味をよく表現していると思われ，かつ意味が明確に異なる extinct と whale のそれぞれを基準列とする．

DMC の実行設定は次のとおりである．初期部分行列は 2 つの基準列に対して同じものにするため $R_l \times C_s$ を元にする．R_s を用いると，extinct を含む文書集合と whale を含む文書集合が異なるため，基準列によって初期行が変わる．C_l を用いても同様で，extinct と同じ文書に出現する語幹集合は whale と同じ文書に出現する語幹集合と異なるため，基準列によって初期

☆ R_{sl} は基準行と共通要素を持たない行の一部を含むが，これらの行はすべてスコア 0 になるため区別できる．

☆☆ http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.html

☆☆☆ http://hltk.sourceforge.net/api-1.3/public/nltk.stemmer.porter-module.html

a / addit / among / analys / and / appear / as / at / atlant / be / biogeograph / bowhead / but / by / carri / characterist / commerci / compar / conserv / context / crb / current / decis / demograph / demographi / detail / determin / differ / discret / distinct / distribut / divers / dna / dr / drove / dure / each / effect / endang / exploit / extinct / facilit / flow / further / gene / genealog / genet / gray / great / ha / hawaii / hemispher / histor / histori / humpback / hundr / in / inform / intern / into / is / isol / it / level / life / magnific / mammal / manag / may / migratori / minim / mitochondri / mysticet / near / northern / ocean / of / on / out / over / own / pacif / palumbi / past / pattern / permit / polici / pool / popul / prior / provid / regard / relationship / size / smaller / somewhat / southern / speci / stephen / structur / studi / subdivid / the / these / thi / three / to / two / univers / variat / whale / wide / will / with / world / year

図 9 a9000006 に含まれる語幹のリスト
Fig. 9 A list of stems included in base-row a9000006.

列が変わる．

$R_l \times C_s$ に対する変更は次のとおりである．R_l ではサイズ制御を行い，まず総行数の 1 割以上の非ゼロ要素を持つ列を無効列として R_l 決定時に無視する．これは，図 8 (b) に示されるように，非ゼロ要素数のきわめて多い列が存在するためである．次に，各行のスコアを決定する．このスコアは，無効列以外で基準行と共通する要素数を x，無効列も含めた非ゼロ要素数を y とすると，x/y である．基準行のスコアは最大値とし，スコアの上位 500 位を初期行とする．初期列は C_s から無効列を排除したものとする．得られた初期部分行列のサイズは行数 503 列数 78 である．

縮退方向決定関数は処理速度を重視した次のものを用いる．ここで用いる行数列数比は列数を行数で割った値である．

(1) 引数として初期部分行列の行数列数比 rc および部分行列 M_{sub} を受け取る．
(2) M_{sub} の行数または列数が 1 なら終了を返す．
(3) M_{sub} の密度が 1 なら終了を返す．
(4) M_{sub} が空行を含めば行を返す．
(5) M_{sub} が空列を含めば列を返す．
(6) M_{sub} の行数列数比が rc 以上なら列を返す．
(7) 行を返す．

副方向調整はノルム 1，類似性尺度はノルム 1 の内積，縮退操作は縮退率 0.5 の割合縮退とする．

この設定で得られた縮退 MC 順序から基準行・列を除いた上位を**図 10** と**図 11** に示す．図中の行は順位，文書番号，タイトルの順に示され，列は順位，語幹の順に示されている．Whale の行順序では 7 位に同順が多いため，1 位だけを示している．図 10 と図 11 は行順序にも列順序にも共通のものがないが，初期部分行列は同じものであるため，これらの結果は基準行が同じでも基準列の意味が明確に異なれば出力も大きく異なることを示している．

基準列が extinct のとき，行はすべてタイトルに CRB を含んでいる．これは Conservation and Restoration Biology の略であり，extinct（extinction の語幹）と関係する語である．これらの行は鯨に限らない生物を対象にした生息地，個体群の行動パターンや多様性に注目した研究に相当するため，出力行は基準行を extinct の観点から見たときの関連文書と見なせる．列順序では popul（population），migratori（migratory），divers（diversity）といった語幹の順位が高くなっているが，これらも基準要素を反映している．

行順序:
1: a9000075: CRB: Migratory Behavior of the Olive Ridley Sea Turtle
1: a9424615: CRB: Projecting Diversity of Neotropical Migratory Birds Under Global Climate Change
1: a9302247: CRB: Metapopulation Structure of the Mexican Spotted Owl: Theory and Conservation Implications
1: a9100397: CRB: Demography and Genetics of Rare Plants
5: a9300182: CRB: Population Viability and Biodiversity Following Rainforest Fragmentation
6: a9000486: CRB: Population Viability of Tropical Forest Vertebrates
6: a9424595: CRB: Should Molecular Genetic Diversity be Used as a Predictor of Evolutionary Potential?
8: a9225081: CRB: Effects of Dispersal on Demography and Genetic Varia- bility in Small Isolated Populations of the Northern Idaho Ground Squirrel: A Model System for Species Threatened...
9: a9000091: CRB: Genetic Variation and Estimates of Population Viability for a Rare Perennial Plant
9: a9025018: CRB: Phenotypic Variation in Megaherbivores
9: a9100002: CRB: Population Size and Density Effects on Population Viability: A Case Study of Two Cirsium Species
9: a9322544: CRB: Mating System and Effective Population Size: An Analysis of Philopatry and Multiple Paternity in Green Sea Turtles
列順序:
1: conserv / 2: crb / 3: mammal / 4: popul / 5: speci / 6: migratori / 7: divers / 8: great / 9: manag / 10: commerci

図 10 Extinct の縮退 MC 順序
Fig. 10 DMC order for “Extinct”.

行順序:
1: a9208369: Public Attitudes to Whales and Whaling: An International Study
1: a9321112: Archaeology of the North Alaska Coast: A Settlement Pattern Study from Point Franklin to Wainwright
1: a9213281: Marine Biotech Fellowship: A Test of Alternative Speciation Hypotheses Using Molecular Systematic Analysis of Lagenorhynchus Dolphins.
1: a9101034: South Channel Ocean Productivity Experiment (SCOPEX): Springtime Circulation in the Great South Channel
1: a9419617: Reconstructing Phylogenetic Frameworks Using Marsupials as a Test System, With Implications For Marsupial Biogeography and the Evolution of Morphological and Molecular Characters
1: a9024592: U.S.-Sweden Cooperative Research: Genetic Diversity of Endangered Populations of Mysticete Whales: Mitol Historical Demographic Studies
列順序:
1: mysticet / 2: stephen / 2: palumbi / 4: humpback / 5: gray / 5: drove / 5: magnific / 5: bowhead / 9: intern / 10: dna

図 11 Whale の縮退 MC 順序
Fig. 11 DMC order for “Whale”.

それに対して基準列が whale のときの行は，基準行と同様の生物学的研究のほかに，捕鯨に関する世界各国の意識調査など，extinct の場合と比べてばらつきがある．列順序は whale との関連が理解しやすく，mysticet，humpback，gray，bowhead といった鯨の種類を示す語幹の順位が高くなっている．そのほかに申請者の名前 stephen palumbi の順位が高くなっているが，これは全文書中で基準行と基準行の研究の拡張と明言されている図中の a9024592 にしか登場しない．2 つの文書にしか登場しない著者名が高い位置にあることは，行に意味上のばらつきがあることを反映していると考えられる．

このときの DMCDR の処理時間は，我々の実装でそれぞれ 0.19 秒と 0.18 秒であった．利用した計算機は CPU が PentiumIV 3.2 GHz，1 GB RAM の PC で，OS は Linux，コンパイラは GCC，プログラミング言語は C++である．

6.2 ベクトル空間法との比較

同じ文書に対してベクトル空間法で決定した行順序の上位 10 位と，extinct と whale を基準列としたときの縮退 MC 順序でそれらの行が占める順位を**図 12** に示す．図中の e は extinct，w は whale の順位を示す☆．ベクトル空間法の類似性尺度は内積のノルム 1 とし，DMCDR の場合と同様に総行数の 1 割以上に非ゼロ要素を持つ列は無視した．

図 12 の行は 2 つの縮退 MC 順序の一方のみで順位が高いものと，両方で順位の低いものに大きく分けられる．Extinct と whale の両方で順位の低いものがあることは，これらのほかに文書間の類似性に強く影響する語幹が存在することを示す．また，a9113342 は初期部分行列から排除されたため，2 つの縮退 MC 順序のいずれにも含まれない．それにもかかわらず順位が高い原因は，基準行文書と a9113342 の両者において無効列以外で最も要素の値の大きい語幹 popul の影響と思われる．

7. おわりに

本論文で我々は行列上の相互順序決定を行うアルゴリズムである DMCDR を定義した．また，実際の文書データにこのアルゴリズムを適用し，妥当な出力が得られることを確認した．

なお，本論文のアルゴリズムはある種の行列を適切に扱えない．その例を**図 13** に示す．(a, A) を基準要素，行列を二部グラフの隣接行列と見なすと，b，c は a と連結，d，e は非連結である．そのため，直感的には b，c と d，e を区別するのが妥当と思われる．しかし，本論文で最大の初期行集合 R_{sl} でも c を含まず，何からの方法で強制的に c を初期行集合に含めても，本論文の類似性尺度は c と a を無関係とする．DMCDR でこのような行列から妥当な出力を得るには，行列に対する何らかの前処理か，R_{sl} や C_{sl} より大きい初期部分行列とそれに対応する類似性尺度が必要である．さらに，与えられた行列にどのような初期部分行列と類似性尺度が適切かを判断する機械的な方法があると有用である．

本論文のアルゴリズムの拡張については次の点が考えられる．我々は文献 3) の BIpDMC を基にしたが，文献 12) はそれを応用上の要求に従って拡張している．その拡張では 1 つの基準行の中に複数の基準列を指定でき，それらの列は基準列の一部の性質のみを持つ準基準列として扱われる．この拡張は本論文の DMC でも可能であると思われ，さらに進めると，基準行が 1 つという制約も取り払い，任意の位置の任意の個数の

ベクトル空間法	e	w
1: a9024592: U.S.-Sweden Cooperative Research: Genetic Diversity of Endangered Populations of Mysticete Whales: Mitol Historical Demographic Studies	46	1
2: a9208369: Public Attitudes to Whales and Whaling: An International Study	468	1
3: a9113342: Evolutionary Ecology of Structured Populations	−	−
4: a9000063: Population Biology of Tropical Rain Forest Trees	13	168
5: a9207278: Doctoral Dissertation Research in Geography and Regional Science	167	168
6: a9212583: Dissertation Research: Importance of Genetic Factors on Fecundity and Survival of Small Populations	65	168
7: a9207558: Likelihood Methods for Population Samples of Sequences	39	145
8: a9307694: Nonadditive Genetic Variance: The Genetical Consequences of Population Structure	139	168
9: a9211945: ABR: Developments in Matrix Population Analysis	18	168
10: a9100002: CRB: Population Size and Density Effects on Population Viability: A Case Study of Two Cirsium Species	9	168

図 12 ベクトル空間法の出力と縮退 MC 順序の比較
Fig. 12 Comparison between vector-space model and DMC order.

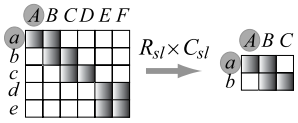


図 13 本論文のアルゴリズムが適切に扱えない行列
Fig. 13 A matrix that can't be treated by this algorithm.

☆ Whale の順位に多い 168 は同時に削除された空行に相当する．空行であるためスコアと順位が等しくなる．

非ゼロ要素を基準要素とすることが考えられる．このようなアルゴリズムの考案は今後の課題である．

参 考 文 献

1) 小柳 滋, 久保田和人, 仲瀬明彦：Matrix Clustering：CRM 向けの新しいデータマイニング手法, 情報処理学会論文誌, Vol.42, No.8, pp.2156–2166 (2001).
2) Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning, *Knowledge Discovery and Data Mining*, pp.269–274 (2001).
3) 上原子正利, 小柳 滋：内積縮退 MC：類似行の検出と類似列の検出を組み合わせたマトリクスクラスタリングアルゴリズム, 情報処理学会論文誌：データベース, Vol.45, No.SIG 7 (TOD22), pp.151–162 (2004).
4) 徳永健伸：情報検索と言語処理, 東京大学出版会 (1999).
5) Agrawal, R., Imielinski, T. and Swami, A.: Mining Association Rules between Sets of Items in Large Databases, *Proc. 1993 ACM SIGMOD International Conference on Management of Data*, pp.207–216 (1993).
6) Agrawal, R. and Srikant, R.: Fast Algorithms for Mining Association Rules, *Proc. 20th VLDB Conference*, pp.487–499 (1994).
7) Brin, S., Motwani, R. and Silverstein, C.: Beyond Market Baskets: Generalizing Association Rules to Correlations, *Proc. ACM SIGMOD Conference on Management of Data*, pp.265–276 (1997).
8) Resnick, P. and Varian, H.R.: Recommender System, *Comm. ACM*, Vol.40, No.3, pp.56–58 (1997).
9) Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl, J.: GroupLens: An Open Architecture for Collaborative Filtering of Netnews, *CSCW94*, pp.175–186 (1994).
10) McKeown, K.R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J.L., Nenkova, A., Sable, C., Schiffman, B. and Sigelman, S.: Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster, *HLT '02* (2002).
11) Mandhani, B., Joshi, S. and Kummamuru, K.: A Matrix Density Based Algorithm to Hierarchically Co-Cluster Documents and Words, *WWW2003* (2003).
12) 上原子正利, 池田貴紀, 浅井一希, 古谷楽人, 内藤裕紀, 小柳 滋：ニュース・ウェブログ記事集約サイトの開発, 電子情報通信学会論文誌, Vol.J88-D-1, No.2, pp.305–315 (2005).

(平成 17 年 3 月 20 日受付)
(平成 17 年 6 月 3 日採録)

(担当編集委員 金子 邦彦)

上原子正利

立命館大学研究員．1997 年京都大学工学部情報工学科卒業．1999 年同大学大学院工学研究科情報工学専攻修士課程修了．

小柳 滋（正会員）

立命館大学教授．1972 年京都大学工学部数理工学科卒業．1977 年同大学大学院工学研究科数理工学専攻博士課程修了．同年（株）東芝入社．2002 年より現職．工学博士．データマイニング, 並列処理, コンピュータアーキテクチャに関する研究に従事．電子情報通信学会, ACM, IEEE-CS 各会員．