主専攻実験
数理モデリングとアルゴリズム

課題1：Webページ解析

1　はじめに

本実験では，Web ページのリンク関係を用いて各ページの人気度を測る方法について学ぶ．Web ページ間の閲覧者の移動を行列で表すことで，Web ページを閲覧している人の移動の問題が行列の固有値問題に帰着する．MATLAB を用いて，この移動を表す行列の生成やその固有値計算などを行う．さらに，そのプログラムを用いて実際の Web ページを解析する．

2　Webページのリンク関係の行列による表現

対象とする Web ページの数を n とし，これらのページには 1 から n の通し番号がふられているとする．n 次の行列 $G = (g_{ij})$ は Web ページのリンク関係を表す行列とする．ここで，ページ j からページ i にリンクがある場合には G の ij 要素を $g_{ij} = 1$ とし，そうでなければ $g_{ij} = 0$ とする．これによって，行列 G の各要素をみることで，どのページからどのページにリンクが張られているかがわかる．

行列 G の各列の要素の和を

$$c_j = \sum_{i=1}^{n} g_{ij},\ j = 1, 2, \ldots, n$$

とする．この値はページ j から他のページへのリンクの数を表している．

あるページから他のページにリンクをたどって移動する確率を p とし，$\delta = (1-p)/n$ はランダムにページを選択する確率とする．ページ j からページ i に移動する確率を a_{ij} とすると，

$$a_{ij} = \begin{cases} p \times \frac{g_{ij}}{c_j} + \delta & c_j \neq 0 \\ \frac{1}{n} & c_j = 0 \end{cases}$$

によって与えられる．これを第 ij 要素とする行列を $A = (a_{ij})$ とする．このような行列 A の最大の固有値は $\lambda = 1$ であることが知られている．

$c_j = 0$ のときは，そのページから他のページへのリンクがないことを表している．このときには，すべてのページへ同じ確率で移動するものとして，第 j 列の各要素を $1/n$ とする．

例として，図 1 のようなリンク構造の Web ページを考える．ページ 1 番からはページ 2 とページ 3 にリンクが張られているため，行列 G の第 1 列の要素のうち，2 番目と 3 番目が 1 となり，他は 0 となる．ページ 2 からはページ 1 のみリンクが張られているため，G の第 2 列は 1 番目の要素のみ 1 となり，他は 0 となる．同様に各ページから他のページへのリン

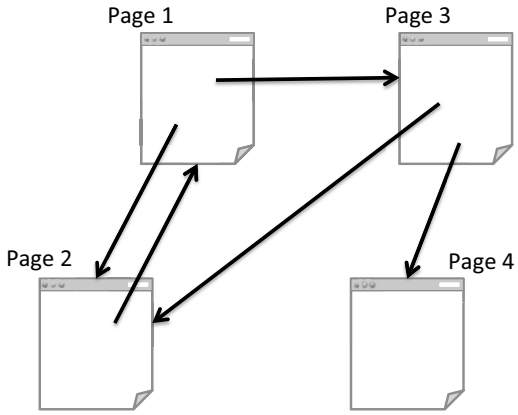

図 1: Web ページのリンク構造の例

クを調べることで，G の各要素の値が決まる．このとき，リンク関係を表す行列 G は

$$G = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

となる．G の各列ベクトルの要素の値の和を要素とする行ベクトルは $\boldsymbol{c} = [2, 1, 2, 0]$ となる．

ここで，$p = 0.8$ とすると，この G に対応する閲覧者のページ間の移動を表す行列 A は

$$A = \begin{bmatrix} 0.05 & 0.85 & 0.05 & 0.25 \\ 0.45 & 0.05 & 0.45 & 0.25 \\ 0.45 & 0.05 & 0.05 & 0.25 \\ 0.05 & 0.05 & 0.45 & 0.25 \end{bmatrix}$$

と表される．

各ページを閲覧している人の割合を要素とする n 次元ベクトルを $\boldsymbol{x}$ とする．このとき，行列 A と $\boldsymbol{x}$ の積 $A\boldsymbol{x}$ は，各ページから 1 回移動をした後の各ページの閲覧者の割合となる．

$A\boldsymbol{x}$ によって $\boldsymbol{x}$ が変化しないとき，すなわち，

$$\boldsymbol{x} = A\boldsymbol{x}$$

となったとき，定常状態になったとみなせる．このとき，ベクトル $\boldsymbol{x}$ の第 j 要素の値は，定常状態になったときにページ j を閲覧している人の割合を表している．

3 MATLAB コードの例

前節の Web ページのリンク構造に対して，G を生成して対応する $\boldsymbol{c}$ を求める MATLAB コード，およびその実行結果を以下に示す．

```
n = 4;
i = [1 2 2 3 4];
j = [2 1 3 1 3];
G = sparse(i,j,ones(1,length(i)),n,n);
G = full(G)
c = sum(G,1)

G =
     0     1     0     0
     1     0     1     0
     1     0     0     0
     0     0     1     0

c =
     2     1     2     0
```

この行列 G をもとにして，移動の確率を表す行列 A を生成する MATLAB コードは以下のように表される．

```
p = 0.8;
d = (1-p)/n;
A = ones(n,n)/n;
for i=1:n
   for j=1:n
      if c(j) ~= 0
         A(i,j) = p*G(i,j)/c(j) + d;
      end
   end
end

A =
    0.0500    0.8500    0.0500    0.2500
    0.4500    0.0500    0.4500    0.2500
    0.4500    0.0500    0.0500    0.2500
    0.0500    0.0500    0.4500    0.2500
```

ベクトル $\boldsymbol{x}$ の要素を

$$\boldsymbol{x} = \begin{bmatrix} 0.6243 \\ 0.5772 \\ 0.4123 \\ 0.3274 \end{bmatrix}$$

と与えたとき，

$$A\boldsymbol{x} = \boldsymbol{x}$$

である．このことから，このベクトル $\boldsymbol{x}$ は固有値 $\lambda = 1$ に対応する固有ベクトルである．

4　固有ベクトルの計算

行列 A に対して固有値と固有ベクトルを MATLAB で求めるには，関数 `eig` を以下のように用いる．

```
[X, D] = eig(A);
lambda = diag(D);
```

ここで，`X` は n 次の正方行列で，各列は固有ベクトルとなっている．また，`D` は n 次で対角要素以外はすべて 0 である．D の対角要素には A の n 個の固有値が並んでおり，その値は `X` の各列のベクトルと対応している．

関数 `diag(D)` は対角行列 `D` の対角要素を取り出す操作を表し，変数 `lambda` はその要素に固有値が並んだ n 次のベクトルとなる．

行列 `X` の第 j 列を取り出す命令は `X(:,j)` である．以下のように

```
A*X(:,1)
```

とすることで，行列 A と行列 `X` の第 1 列のベクトル $\boldsymbol{x}_1$ との積 $A\boldsymbol{x}_1$ の計算を行うことができる．λ_1 が固有値，$\boldsymbol{x}_1$ が固有ベクトルのとき，

```
norm(A*X(:,1) - lambda(1)*X(:,1))
```

によって，残差 $\|A\boldsymbol{x}_1 - \lambda_1\boldsymbol{x}_1\|$ が得られる．

5　実験課題

課題 1-1：Web リンク構造の行列表現

これまでの説明に従って，図 1 に示すようなリンク構造の Web について，リンク関係を表す行列 G，および，それに対応する移動確率を表す行列 A を生成せよ．すべての要素が 1/4 の 4 次元ベクトル $\boldsymbol{x}$ を与えて $A\boldsymbol{x}$ を計算し，その要素がどのように変化するか確認せよ．さらに，$A^2\boldsymbol{x}$，$A^3\boldsymbol{x}$ についても，計算結果のベクトルの要素の値を確認せよ．

課題 1-2：固有値問題解法の利用

A の固有値と固有ベクトルを求めよ．A の固有ベクトルのうち，固有値 1 に対応するベクトルを `y` とし，`y` の要素が大きいものから順に並べ替え，並べ替えた順番を示せ．要素を並べ替える MATLAB の関数は `sort` である．引数などの関数の使い方はコマンドウインドウ上で `help sort` と入力することで得られる．

課題 1-3：簡単な例による Web ページ人気度の計算

図 2 に示すようなリンク関係を表す行列 G，および移動確率を表す行列 A を求めよ．この A を用いてページの人気度を計算し，人気度順にページ番号を示せ．

課題 1-4：実際の Web による解析

MATLAB の関数 `surfer` は，与えられたページから下のリンク構造を返す．適当な Web サイトを選び，この関数を用いてリンク構造を求め，ページの人気度を解析せよ．その結果

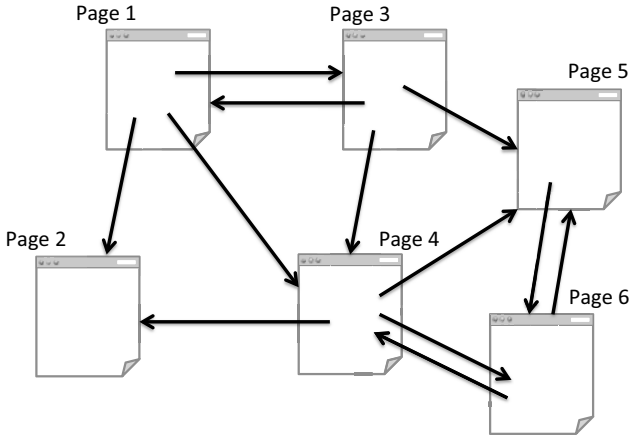

図 2: Web ページのリンク関係の例 2

について考察し，この方法によって得られる結果に問題点がないかどうか検討せよ．もし，問題点が見つかった場合，それに対してどういう解決策が考えられるか述べよ．

関数 `surfer` は

```
root = 'http://www.......';
m = 10;
[U,G1] = surfer(root,m)
G = G1 - diag(diag(G1));
G = full(G);
[n,n] = size(G);
```

のように用いる．引数の `root` には対象とする Web ページ群のトップページの URL を与え，そのページ中のリンクから始めて `m` 個のリンクをたどる．戻り値 `U` にはリンクをたどった URL のリストが入り，`G1` によってリンク構造を表す行列が得られる．ただし，`G1` にはページ内へのリンクも含まれるため（対角要素が 1 になる），それを取り除いて `G` が得られる．このとき，`G` は 0 でない要素のみのデータだけを保持する疎行列のデータ形式であるため，関数 `full(G)` によってすべての要素のデータを保持する形式に変換する．

`surfer.m` は https://jp.mathworks.com/matlabcentral/fileexchange/37976-numerical-computing-with-matlab よりダウンロードすること．

参考文献

Cleve Moler, Numerical Computing with MATLAB, `http://www.mathworks.co.jp/moler/index ncm.html`