マルコフ連鎖の極限推移確率とWebリンク解析

岡村　寛之

マルコフ連鎖の推移確率に関する極限定理および関連する極限分布と定常分布の関係はマルコフ連鎖において非常に重要である．それ故，マルコフ連鎖を扱う教科書では必ずといって良いほど，定常分布や極限分布の導出に対する例題がある．ここでは，マルコフ連鎖の極限確率に関する例題を示した後に，Webリンク解析として知られているPageRankアルゴリズムが本質的に教科書に記されている定常確率に関する例題と同様な解析であることを紹介する．

キーワード：離散時間マルコフ連鎖，極限確率，定常確率，PageRank

1. はじめに

マルコフ連鎖とは，未来のふるまいが現在の状態だけで決定され，過去のふるまいに依存しない性質（マルコフ性）を有する確率過程の一種であり，待ち行列や信頼性など確率モデルを用いた性能評価に応用されている．

マルコフ連鎖を用いた確率モデルで性能評価を行う場合，マルコフ連鎖の推移確率に関する極限定理および関連する極限分布と定常分布の関係を理解することは非常に重要である．それ故，マルコフ連鎖を扱う教科書では必ずといって良いほど，定常分布や極限分布の導出に対する例題がある[1]～[3]．しかしながら，極限分布と定常確率の数学的な議論が現実の問題とどのように関連しているのかを理解するのは，初学者にとってなかなか難しい問題である．本稿では，マルコフ連鎖の極限確率や定常分布の解析の一例として，Webリンク解析で知られているPageRankアルゴリズムを紹介し，PageRankが本質的に教科書に記されている定常確率に関する例題との同様な解析であることを紹介する．

Webリンク解析はWebに散在するドキュメント（HTML）の関連性を定量化することを目指しており，Webの情報検索における大きな課題である．伝統的にドキュメント間の類似度などによる関連付けは，ドキュメント内のキーワードによる手法が行われてきている．一方で，WebにおけるHTMLの大きな特徴の一つは他のドキュメントに対するリンクをつなぐハイパーリンクと呼ばれる構造であり，そのリンクはドキュメント作成者が「重要」あるいは「関連がある」と認識しているドキュメントに張られる．換言すると，リンク構造そのものが何らかの情報を有しており，実際に1990年代末ごろから積極的にリンク情報がWeb検索に用いられている．

相互関係を表すネットワークに関しては，社会科学の分野においても精力的に研究されている．例えば，ネットワーク形成については，Erdos and Rényi[4]によるランダムネットワーク，Watts and Strogatz[5]によるスモールワールドネットワーク，Barabáshi and Albert[6]によるスケールフリーネットワークなどがある．他方，Webリンク解析と同様なアプローチとして，論文の引用・被引用の関係から雑誌の水準を評価するインパクトファクタがある[7]．

一方，ハイパーテキストのリンク構造を情報検索に利用することは医学の分野などで関心を持たれ[8]，初期のWebリンク解析に関する研究はBray[9]やMarchiori[10]で行われている．また，Kleinberg[11]はWebリンク解析に対する代表的な手法の一つであるHITS（Hypertext Induced Topic Selection）を提案している．Webリンク解析が従来の相互関係ネットワークの解析と異なるのは，ページおよびリンク数のスケールであり，Webリンク解析では，高いスケーラビリティを有したアルゴリズムであることが望まれる．その点でHITSは何らかの手段で選ばれたハイパーテキストの集合に適用される手法であり，Web全体のリンク構造を解析するには至っていない．

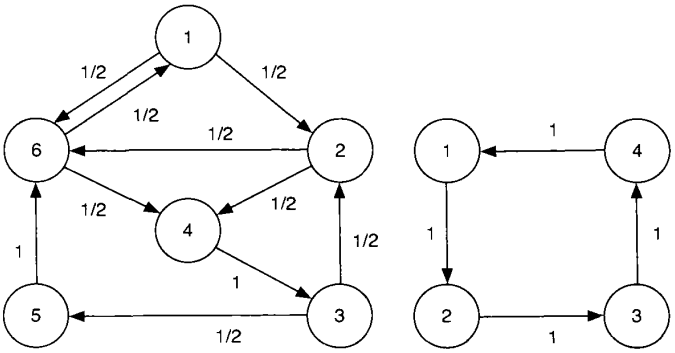
おかむら　ひろゆき
広島大学 大学院工学研究科
〒739-8527 東広島市鏡山1-4-1

Pageら[12][13]はPageRankと呼ばれるWebリンク構造を解析して各ページをスコアリングする手法を提案した．これはHITSよりもシンプルなアルゴリズムであり，このスコアリングアルゴリズムを適用したWeb検索サイトが現在のGoogleの原型となっている．

一見して，PageRankとマルコフ連鎖はまったく関係がないように思われるが，実際はWebを閲覧する行動を確率的に表したモデル上で，教科書に掲載されているようなシンプルな手法で極限確率や定常分布を求めた結果として得られる．以降では，離散時間マルコフ連鎖の例題を示し，極限分布と定常分布の議論を解説した後に，PageRankを導出するための確率モデルとPageRankアルゴリズム，およびマルコフ連鎖の極限分布と定常分布の関連について説明する．

2. 離散時間マルコフ連鎖

例題：次のマルコフ連鎖の極限分布と定常分布を求めよ．



2.1 定義

上の例題を解くために必要な，離散時間マルコフ連鎖の極限分布と定常分布に対する基礎的な解説を与える（定義および定理の詳細は例えば文献[1]～[3]を参照）．

離散時間の確率過程$\{X_n;\ n=0,1,2,\cdots\}$で状態空間Sが有限かつ離散的なものを考える．一般性を失うことなく$S=\{1,2,\cdots,k\}$とする．このとき，すべての$n\geq 0$とすべての$j\in S$に対して

$$P(X_{n+1}=j|X_0=i_0,X_1=i_1,\cdots,X_n=i)=P(X_{n+1}=j|X_n=i) \tag{1}$$

が成り立つとき，確率過程$\{X_n;\ n=0,1,2,\cdots\}$は離散時間マルコフ連鎖と呼ばれる．上記の確率がnと無関係になる場合（斉次マルコフ連鎖）を考え

$$p_{i,j}=P(X_{n+1}=j|X_n=i) \tag{2}$$

とおくと，$p_{i,j}$は離散時間マルコフ連鎖の推移確率と呼ばれ，推移確率$p_{i,j}$を要素とする行列$\boldsymbol{P}=[p_{i,j}]$は推移確率行列として定義される．推移確率行列は各行の要素の和が$\sum_{j=1}^{k}p_{i,j}=1$となる非負の要素からなる行列で，このような行列$\boldsymbol{P}$を一般に確率行列と呼ぶ．

いま，$n=0$における初期分布

$$\boldsymbol{\pi}(0)=(\pi_1(0),\cdots,\pi_k(0)), \tag{3}$$

$$\pi_i(0)=P(X(0)=i),\quad i=1,\cdots,k \tag{4}$$

を定義する．このとき，任意の$n\geq 0$に対して，nステップで状態iから状態jに推移するnステップ状態推移確率を

$$p_{i,j}^{(n)}=P(X(n+m)=j|X(m)=i)=P(X(n)=i|X(0)=i) \tag{5}$$

とすると，任意の$r(0\leq r\leq n)$に対して

$$p_{i,j}^{(n)}=\sum_{l=1}^{k}p_{i,l}^{(r)}p_{l,j}^{(n-r)} \tag{6}$$

が成り立つ（チャップマン・コルモゴロフ方程式）．さらに上記のnステップ状態推移確率に対する行列表現$\boldsymbol{P}^{(n)}=[p_{i,j}^{(n)}]$を用いると，

$$\boldsymbol{P}^{(n)}=\boldsymbol{P}^{(r)}\boldsymbol{P}^{(n-r)}=\boldsymbol{P}^{(1)}\boldsymbol{P}^{(n-1)}=\boldsymbol{P}\boldsymbol{P}^{(n-1)}=\boldsymbol{P}^{n} \tag{7}$$

となる．これより，離散時間マルコフ連鎖におけるnステップ後の状態確率ベクトルは

$$\boldsymbol{\pi}(n)=(\pi_1(n),\cdots,\pi_k(n))=\boldsymbol{\pi}(0)\boldsymbol{P}^n,\quad \pi_i(n)=P(X(n)=i),\quad i=1,\cdots,k \tag{8}$$

で算出できる．

2.2 状態の分類

離散時間マルコフ連鎖のnステップ状態推移確率$p_{i,j}^{(n)}$が非負となる$n\geq 0$が存在するとき，状態iは状態jへ到達可能といい，$i\to j$と書く．状態iと状態jが$i\to j$かつ$j\to i$であるとき，これらは相互に到達可能であるといい$i\leftrightarrow j$と書き，相互到達可能の関係は同値関係である（反射律，対称律，推移律が成り立つ）．この関係によって，マルコフ連鎖のすべての状態Sを同値類に分けることができる．さらに，マルコフ連鎖のすべての状態Sがただ1つの同値類になる場合，このマルコフ連鎖は既約であるという．

マルコフ連鎖の状態iの周期$d(i)$は，$p_{i,i}^{(n)}>0$となる$n\geq 0$の最大公約数で定義される．周期が$d(i)=1$のとき，状態iは非周期的という．また，同じ同値類に属する状態は同じ周期になる．

離散時間マルコフ連鎖に対して，状態iから出発してnステップで初めて状態jに到達する初到達確率を次のように定義する．

$$f_{i,j}^{(n)}=P(X(n)=j,\ X(r)\neq j,\ r=1,\cdots,n-1|X(0)=i). \quad (9)$$

このとき，状態 i から状態 j へいつか到達する確率は

$$f_{i,j}=\sum_{n=1}^{\infty}f_{i,j}^{(n)}. \quad (10)$$

いま，$f_{i,i}=1$ であるならば，状態 i は再帰的であるといい，$f_{i,j}<1$ であるならば，状態 i は一時的あるいは過渡的という．

2.3 極限分布と定常分布

状態 i から出発して状態 j に初めて到達するまでの期待ステップ数を $\mu_{i,j}$ とすると，初到達確率を用いて

$$\mu_{i,j}=\sum_{n=1}^{\infty}nf_{i,j}^{(n)} \quad (11)$$

と表せる．このとき，$\mu_{i,i}<\infty$ であるならば状態 i は正再帰的，$\mu_{i,i}=\infty$ であるならば状態 i は零再帰的と呼ばれる．状態 i が一時的あるいは過渡的であるならば明らかに $\mu_{i,i}=\infty$ となる．一方で，状態 i が再帰的であっても $\mu_{i,i}$ が無限大になることもある（状態数が有限のマルコフ連鎖では零再帰的状態は存在しない）．

以上の準備のもとで，離散時間マルコフ連鎖の極限に関する以下の定理を与える．

定理 1：既約で正再帰的なマルコフ連鎖において，連鎖が非周期的ならば極限における推移確率は

$$\lim_{n\to n}p_{i,j}^{(n)}=\frac{1}{\mu_{j,j}},\quad i,j=1,\cdots,k \quad (12)$$

のように初期状態 i と独立となり，$n\to\infty$ における状態確率（極限分布）は

$$P(X(\infty)=j)=1/\mu_{j,j},\quad j=1,\cdots,k \quad (13)$$

で与えられる．一方，連鎖の周期が $d\geq 2$ ならば，$p_{i,j}^{(n)}$ は収束せず，極限分布も存在しない．

定理 2：既約で正再帰的なマルコフ連鎖において極限分布は，ベクトル $\boldsymbol{\pi}=(\pi_1,\cdots,\pi_k)$ に対する以下の平衡方程式の解として得られる．

$$\boldsymbol{\pi}=\boldsymbol{\pi}P,\quad \sum_{i=1}^{k}\pi_i=1. \quad (14)$$

このベクトル $\boldsymbol{\pi}$ はマルコフ連鎖の定常分布と呼ばれる．

上述の定理は，マルコフ連鎖に極限分布が存在するための必要十分条件が，正再帰的かつ非周期的であり，平衡方程式の解（定常分布）が極限分布に一致することを示している．

解答：有限状態の離散時間マルコフ連鎖(a)は既約かつ非周期的なので，極限分布と定常分布が同じになる．いま $\boldsymbol{\pi}$ を定常分布とすると平衡方程式

$$\boldsymbol{\pi}=\boldsymbol{\pi}\boldsymbol{P}_a,\quad \sum_{i=1}^{6}\pi_i=1$$

を満たす．ここで $\boldsymbol{P}_a$ は(a)のマルコフ連鎖から作られる推移確率行列で

$$\boldsymbol{P}_a=\begin{pmatrix}0 & 1/2 & 0 & 0 & 0 & 1/2\\ 0 & 0 & 0 & 1/2 & 0 & 1/2\\ 0 & 1/2 & 0 & 0 & 1/2 & 0\\ 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 1\\ 1/2 & 0 & 0 & 1/2 & 0 & 0\end{pmatrix}$$

となる．これを解くと，極限分布および定常分布として

$$\boldsymbol{\pi}=(\pi_1,\pi_2,\pi_3,\pi_4,\pi_5,\pi_6)=\left(\frac{6}{51},\frac{8}{51},\frac{10}{51},\frac{10}{51},\frac{5}{51},\frac{12}{51}\right)$$

が得られる．

有限状態の離散時間マルコフ連鎖(b)は既約であるが周期的（周期 4）であるので，極限分布が存在しない．一方，定常分布は次の平衡方程式の解として得られる．

$$\boldsymbol{\pi}=\boldsymbol{\pi}\boldsymbol{P}_b,\quad \sum_{i=1}^{4}\pi_i=1.$$

ここで $\boldsymbol{P}_b$ は

$$\boldsymbol{P}_b=\begin{pmatrix}0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 1 & 0 & 0 & 0\end{pmatrix}$$

である．これより，定常分布として

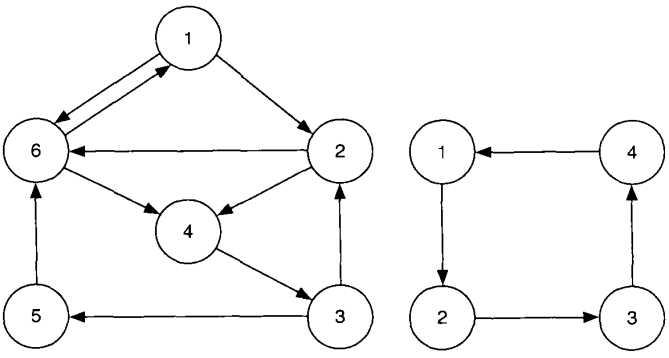
$$\boldsymbol{\pi}=(\pi_1,\pi_2,\pi_3,\pi_4)=\left(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right)$$

を得る．

---

3. Web リンク解析

例題：次の Web リンク構造があったとき各ページの重要度を算出せよ．ただし，以下の Web グラフ上のノードは HTML，アークは HTML 間のリンク（リンク元→リンク先）を示している．



3.1 PageRankアルゴリズム

PageRankはPageら[12][13]によって提案されたスコアリングであり，検索サイトGoogleにおけるスコアリングアルゴリズムとして知られている．PageRankは質問非依存のアルゴリズムであり，クエリに関係なく算出される値であり，この点もPageRankの高いスケーラビリティに貢献している．PageRankは「良質なページからリンクされるページも良質なページである」という基本的な考えに基づいてページごとの重要度が算出される．PageRankアルゴリズムの説明では，「支持投票」や「行列固有値計算」に関する記述が数多く見受けられるが，ここではPageRankを確率モデルの視点から解説する．

PageRankアルゴリズムはWeb上を巡回するユーザのふるまいをモデル化することで説明される（ランダムサーファーモデルと呼ばれる）．いま，1人のユーザがWeb上を巡回するモデルを考える．ユーザによるWebの巡回は，あるページからリンクをたどって別のページへ移動することに対応する．ユーザのページ移動に関して次の仮定を設ける．

- 次のページへの移動は確率的である．
- 次のページへの移動は現在訪れているページのみに依存し，過去に訪れたページに依存しない．

いま，Web全体のページを$S=\{1,2,\cdots,k\}$とラベル付けし，確率過程$\{X(n);\ n=0,1,\cdots\}$をn回のページ移動をしたときにユーザが訪問しているページを表す確率過程としたとき，上述の仮定から，Web上を巡回するユーザのふるまいを記述する確率過程$X(n)$はS上で定義された離散時間マルコフ連鎖となる．

このモデルに従うと，ユーザが良質のページをリンクをたどって移動するとき，ユーザが訪問する確率の高いページがより重要度が高いことがわかる．つまり，PageRankはマルコフ連鎖でモデル化されたWeb上を巡回するユーザが無限回の移動を行ったときに，どのページを訪問しているかを表す確率として定義される．これは，先に解説した離散時間マルコフ連鎖に関する極限分布を求める問題に帰着される．

より詳細な定式化では次のようになる．ページiがリンクを張っているページ（リンク先）の集合を$S_i(\subseteq S)$，そのページ数を$|S_i|$として表す．ユーザがページiを訪問しているとき，他のページ$j(j\neq i)$への移動は確率的であり，その確率を$p_{i,j}$とする．さらに$p_{i,j}$を(i,j)要素とする推移確率行列を$\boldsymbol{P}$とする．

移動に関する確率法則として，PageRankではリンク先へ等確率で移動する戦略を考えている．すなわち，ページiからページjへリンクが張られているならば，推移確率行列$\boldsymbol{P}$の(i,j)要素は$p_{i,j}=1/|S_i|$となり，リンクがない場合は$p_{i,j}=0$となる．

問題は推移確率行列$\boldsymbol{P}$に対する極限分布を求めることであるが，先に示したように，極限分布が存在するためには対象とするマルコフ連鎖が既約かつ正再帰的で非周期でなければならない（有限なマルコフ連鎖であるため実際には正再帰的の条件は除くことができる）．しかしながら，一般のWebリンク構造を扱う場合，リンク構造が既約であることはなく，リンクの存在しないページ（吸収状態）やリンクの張られていないページ（過渡状態）が数多く存在する．また，マルコフ連鎖の周期については膨大なWebリンクデータに対して周期のチェックをすることは困難である．これを回避するために，PageRankのアルゴリズムではリンクをたどって移動する以外に，確率εでユーザが任意のページに（リンクをたどらず）ジャンプする可能性を考慮している．このジャンプはすべてのページに均等に訪問する機会を与えるため，本質的にスコアリングを変化させない（ただしジャンプ確率が高いとスコアの差がなくなる）．また，ジャンプを伴うユーザのふるまいは既約かつ非周期的なマルコフ連鎖を保証する．

最終的にPageRankのアルゴリズムは以下の推移確率行列$\boldsymbol{P}'$を持つ離散時間マルコフ連鎖の極限分布として与えられる．

$$\boldsymbol{P}'=\frac{\varepsilon}{|S|}\boldsymbol{E}+(1-\varepsilon)\boldsymbol{P}. \tag{15}$$

ここで$\boldsymbol{E}$はすべての要素が1の行列である．いま，$\boldsymbol{P}'$は既約かつ非周期的であるため，極限分布は次の平衡方程式の解（定常分布）となる．

$$\boldsymbol{\pi}=\boldsymbol{\pi}\boldsymbol{P}',\quad \sum_{i\in S}\pi_i=1. \tag{16}$$

解答：PageRankによる重要度の算出を行う．Webグラフ(a)の各リンクに均等な重みを持たせることで以下の確率行列を得る．

$$\boldsymbol{P}_a=\begin{pmatrix} 0 & 1/2 & 0 & 0 & 0 & 1/2\\ 0 & 0 & 0 & 1/2 & 0 & 1/2\\ 0 & 1/2 & 0 & 0 & 1/2 & 0\\ 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 1\\ 1/2 & 0 & 0 & 1/2 & 0 & 0 \end{pmatrix}.$$

このとき，各ページのPageRankは以下の方程式を

満たす $\boldsymbol{\pi}$ で求まる．

$$\boldsymbol{\pi}=\boldsymbol{\pi}\left(\frac{\varepsilon}{6}\boldsymbol{E}+(1-\varepsilon)\boldsymbol{P}_a\right),\quad \sum_{i=1}^{6}\pi_i=1.$$

ここで $\boldsymbol{E}$ はすべての要素が1の行列である（相対的な順位が必要なだけなので実際には正規化 $\sum_i\boldsymbol{\pi}_i=1$ は必要ない）．ジャンプ確率を $\varepsilon=0.2$ とするとき，各ページの PageRank は次のように得られる．

$$\begin{aligned}\boldsymbol{\pi}&=(\pi_1,\pi_2,\pi_3,\pi_4,\pi_5,\pi_6)\\&=(0.1266,0.1580,0.1852,0.1898,\\&\quad 0.1074,0.2331).\end{aligned}$$

この結果より，各ページは重要度の高い順に 6，4，3，2，1，5 と順位付けられる．

Web グラフ(b)も同様に，リンクから推移確率行列を考えると

$$\boldsymbol{P}_b=\begin{pmatrix}0&1&0&0\\0&0&1&0\\0&0&0&1\\1&0&0&0\end{pmatrix}.$$

ゆえに，各ページの PageRank は以下の方程式を満たす $\boldsymbol{\pi}$ で求まる．

$$\boldsymbol{\pi}=\boldsymbol{\pi}\left(\frac{\varepsilon}{4}\boldsymbol{E}+(1-\varepsilon)\boldsymbol{P}_b\right),\quad \sum_{i=1}^{4}\pi_i=1.$$

これを $\varepsilon=0.2$ のもとで解くと，各ページの重要度として

$$\boldsymbol{\pi}=(\pi_1,\pi_2,\pi_3,\pi_4)=\left(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right)$$

が得られ，すべてのページが同じ重要度であることがわかる．

4. まとめ

本稿では役に立つ例題として，離散時間マルコフ連鎖の定常分布に関する問題を取り上げ，Web リンク解析として有名な PageRank との相違について解説した．離散時間マルコフ連鎖の定常分布の導出は，基礎的な確率の教科書では必ずといっても良いほど取り上げられる問題である．一方で，PageRank は言わずと知れた Google のスコアリングアルゴリズムであり，その解析の基本原理は教科書に載っている定常分布の例題と大きな違いはない．しかしながら，RageRank ではユーザのふるまいをモデル化し直感的にスコアの妥当性を説明している点や，実際的な Web リンクを扱う際に起こる既約性や周期性の問題をシンプルに解決する手法の提供している点など，いわゆる「目の付け所がよい点」が多数存在する．そして何より Google および PageRank を世界的に知らしめたのはその圧倒的なデータ量であり，数億ページからのリンク情報を解析している．教科書の例題にある解析でも，そのスケールを大きくすることで価値ある情報が得られる好例である．

参考文献

[1] 森村英典, 高橋幸雄, マルコフ解析, 日科技連, 1979.

[2] 尾崎俊治, 確率モデル入門, 朝倉書店, 1996.

[3] 伏見正則, 確率と確率過程, 朝倉書店, 2004.

[4] P. Erdos and A. Rényi, On random graphs, *Publ. Math. Debrecen*, 6: 290-291, 1959.

[5] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, *Nature*, 393: 440-442, 1998.

[6] A. L. Barabási and R. Albert, Emergence of scaling in random networks, *Science*, 286: 509-512, 1999.

[7] E. Garfield, Citation indexes for science: a new dimension in documentation through association of ideas, *Science*, 122: 108-111, 1955.

[8] E. F. Mark, Searching for information in a hypertext medical handbook, *Communications of the ACM*, 31: 880-886, 1988.

[9] T. Bray, Measuring the Web, *Proc. 5th Int. Conf. on the World Wide Web*, 993-1005, 1996.

[10] M. Marchiori, The quest for correct information on the Web: hyper search engines, *Proc. 6th Int. World-Wide Web Conf.*, 1225-1235, 1997.

[11] J. Kleinberg, Authoritative sources in a hyperlinked environment, *Proc. 9th Ann. ACM-SIAM Symp. on Discrete Algorithms*, 668-677, 1998.

[12] L. Page, S. Brin, R. Motwani and T. Winograd, The PageRank citation ranking: bringing order to the Web, Technical report, Stanford University, 1998.

[13] S. Brin and L. Page, The anatomy of a large-scale hypertextual (Web) search engine, *Proc. 7th Int. World-Wide Web Conf.*, 107-117, 1998.