# ウェブページのランキング技術

宇野　裕之

検索エンジンがウェブページにその重要度を与えるランキングの基礎技術について解説する．とくに，ウェブページのリンク構造に基づき，現在の高性能な検索エンジンのランキング手法の先駆や原型となっているHITS と PageRank のアイデアを紹介する．そのうえで，いくつかの発展的な話題を取り上げ，ランキング技術の現状を理解する．

キーワード：ウェブグラフ，ウェブアルゴリズム，行列固有値計算，ハブ-オーソリティ・モデル，マルコフ連鎖，ランダムサーファー・モデル，リンク解析，HITS，PageRank．

## 1. はじめに

ウェブは，いまや多くの人にとってテレビや新聞と並ぶ，あるいはそれらをも包括しうる日常の情報源となった．そこに存在する膨大かつ無秩序な情報を整理し提示する目次や索引の役割りを果たすのが検索エンジンである．ウェブ上の必要な情報を検索エンジンなしに発見することは実質的には困難であり，検索エンジンはもはやウェブにアクセスするために誰もが無意識に利用する不可欠なインフラとなっている．

そのような検索エンジンは，コンピュータサイエンス分野の技術の顕著な成果であり象徴的な成功例の 1 つである．そしてその中には，ハードウェアからソフトウェアまでさまざまな技術のエッセンスが詰まっている．本稿では，その中でも検索エンジンの技術の中核をなす，ウェブページにその重要度を与えるランキングの技術について解説する．とくに，ウェブページのリンク構造に基づき，現在の高性能な検索エンジンのランキング手法の先駆や原型となっている HITS と PageRank のアイデアを紹介する．

## 2. 検索エンジンの要件

ユーザは通常，知りたいことや調べたい事柄に関連する**検索キーワード**を入力することで，検索エンジンに対する**クエリ**を発する．このとき検索エンジンの使命は，「クエリにおけるユーザの意図を正確に把握し，そのニーズにぴったり一致する結果を返す」[11] ことに尽きる．あるはずの探していた情報が見つかるだけでなく，予期しない有益な情報が得られれば，検索エンジンの価値はさらに増す．この使命を果たすために検索結果に求められる要件は，その重要度（関連性），網羅性，リアルタイム性，応答速度などとされている [1][4][11][21]．

まず，検索結果にはユーザが求める情報をもつウェブページが含まれていなければならない．そのためには検索エンジンがそのページを発見して保持していなければ，それを結果として提示することができない．すなわち検索エンジンには，ウェブ上に存在するページをできる限り多く収集している**網羅性**が求められる．そのため検索エンジンは，ウェブページを収集して異なる各ページに番号を与えるインデックス化を行うために，ページのリンクをたどり続けて新しいページを発見するクローラと呼ばれる自動プログラムを動かしている．高性能なクローラによりいかに多くのページをインデックス化しているかは，検索エンジンの性能の指標の 1 つである．

さらに，検索結果には**リアルタイム性**（情報の新しさ）が求められる．インデックス化されたページがいかに多くとも，そのページ内容が更新されたりページが消滅するなどして情報が失われていれば，検索結果は役に立たない．リアルタイム性を維持するためには，クローラの効率的な動作，とくに更新頻度が高い重要なページにはより頻繁に巡回させる技術が必要になる．

そして検索結果は，**重要度**あるいはクエリに対する**関連性**の高い順に提示されることが不可欠である．なぜなら，ユーザの大部分は検索結果表示の 1〜2 ページ目（上位 20 件程度）までしか見ないと言われており，それより下位の結果は参照されなければ情報としての価値がない．ウェブページの重要度（関連性）を算出する根拠となるページごとの得点を**スコア**，そのスコアにしたがって与えられる順位を**ランク**といい，これ

うの　ゆうし
大阪府立大学 理学系研究科 情報数理科学専攻
〒 599–8531 大阪府堺市中区学園町 1–1

  

ら算定の作業あるいは結果をウェブページの**ランキング**と総称する．

検索エンジンにとってこれらいずれの要件も欠くことはできないが，なかでもランキングの良し悪しは検索エンジンの性能を決定づける重要な要素になっている．また，技術的にも工夫の可能性が大きく，検索エンジンが中核の技術としてその独自性や性能を発揮できる舞台となっている．

## 3. 初期の検索エンジンとスパム

ウェブ（World Wide Web）は，1990 年に Tim Berners-Lee によって発明された．その後まもなく出現した archie という FTP サイトのファイル検索アプリケーションは，時代と用途から利用者は限られたが最初の検索アプリケーションの 1 つと言える．1993 年に Mosaic というブラウザが誕生し，多くのユーザがウェブを利用し始めその楽しさに触れることになった．ウェブページが増加するとともにそれを整理する必要性も高まった．1994 年に設立された Yahoo!が用いた仕組みはディレクトリ型検索と呼ばれ，ユーザはその整理された良質な情報に満足したが，技術的には人力の分類による巨大なリンク集に過ぎなかった [2]．

1995 年に設立された AltaVista は，ウェブページの内容を文字列照合などに基づき自動で分類するコンテンツ型検索を採用し，良好な検索結果でユーザに支持され一時代を築いた[1]．技術的には，検索キーワードの出現回数や出現位置，キーワードどうしの近接度，タグ内文字列などのページ内要素でページの関連性を判定しランキングしていた．この方法は，クエリが含む検索キーワードによってページの関連性をその都度計算し，キーワードによって関連性が変化するクエリ従属なランキングであった [2][21]．

ウェブはその普及とともに企業などの商用，宣伝目的での利用が拡大した．例えば，コーヒーショップが“おいしい豆 ケニア” というキーワードによる検索でヒットし，ショップのウェブページが検索結果の上位に表示されるかは，やがてショップの死活問題ともなった．すると，検索エンジンを欺きランクを恣意的に操作する**スパム**行為が横行し始めた．その手法は，ページ内容に関するキーワードをページに何度も埋め込んだり，内容とはまったく無関係な流行語をページに挿入するという稚拙なものであったが，コンテンツ型の検索エンジンにはそのようなスパムに対する耐性はなく，検索結果はひとたまりもなくスパムに荒らされ，検索エンジンのスパムに対する耐性の重要さが認識された．

このような状況の中，1998 年に設立された Google は，より客観的で公平なランキングを求め，ページ外要素を用いるランキング技術を開発した．その技術はスパム問題をも克服し，その検索結果のこれまでにない正確性から急速にユーザの支持を集め，2000 年にインデックス化数（10 億ページ）でナンバーワンを宣言して以降，後発の検索エンジンの挑戦をことごとく退けている[2]．

## 4. リンク解析に基づくページランキング

ページ外要素を利用したウェブページのランキング技法は，2 つのアイデアがほぼ同時期（1998 年）に独立に考案された．1 つが Jon M. Kleinberg による HITS [14] であり，もう 1 つが Larry Page と Sergey Brin による PageRank [27] である．共通するのは，ともにランクを評価するページ外要素としてウェブページ間のハイパーリンクに着目したことである．また，学術文献の相互引用関係にその着想を得ていたことも共通している．この画期的な発想は，停滞していた検索エンジンのランキング技法のブレークスルーとなり，後に‘リンク解析’と呼ばれる研究分野の萌芽ともなった．

ウェブのハイパーリンク構造は，ウェブページを頂点，ウェブページ間に張られたハイパーリンクを有向辺とする有向グラフとみなすことができる．これを**ウェブグラフ**といい，ウェブ上で動作する検索エンジンやクローラなどのウェブアルゴリズム設計のための最も基本的なモデルである．ウェブグラフは構造的にも興味深い性質を数多くもつ [5][15][29]．例えば，ウェブページやリンクの生成死滅にともない，ウェブグラフは動的に変化し成長する．その過程はランダムにも見えるが，古典的なランダムグラフとは大きく性質が異なる．ウェブグラフをその典型例とするスケールフリー・ネットワークやソーシャル・ネットワークは，大きな研究分野を築いている [20][22]．

本節では，HITS と PageRank のアルゴリズムを説

[1] 当時の日本国産の検索エンジンとしては Hole-in-One，千里眼，goo などがあり，とくに goo の日本語に特化した独創的な技術は高く評価された．

[2] ウェブページのランキングの最新技術を Google 抜きに語ることはできないが，それらが高度な企業秘密となった現在，Google の技術を知り語ることもまた難しいというジレンマを抱える．検索エンジン市場の寡占は進み，検索エンジンで検索することを意味する「ググる」という造語も出現した．また，Google の寡占を揶揄して，モノポリーというゲームを模倣した Googolopoly や GooglePoly という非売のゲームも出回った．興味がある方はググってください．

明する．その際，ウェブグラフ $G=(V,E)$ に対してその頂点数を $n\,(=|V|)$ で表し，$N^+(v)=\{w\mid(v,w)\in E\}$, $N^-(v)=\{u\mid(u,v)\in E\}$ と定義する．すなわち，それぞれページ $v$ がリンクを張るページ，ページ $v$ へリンクを張るページの集合を表す．またグラフ $G$ の隣接行列 $A=(a_{uv})$ とは，「$a_{uv}=1 \iff (u,v)\in E$」を満たす $n\times n$ 正方行列である．

### 4.1 HITS

Kleinberg による HITS（hyperlink induced topic search）[14] は，ウェブ上のトピックに関する**ハブ-オーソリティ・モデル**と呼ばれる仮説から始まる．すなわち1つのトピックは，それに関して権威的なページ（オーソリティ）とポータル的なページ（ハブ）それぞれの集合を部集合とする（単に部分グラフとしての）有向2部グラフを構成しているというものである．例えば，サッカーというトピックに関するページの中には，サッカーチームの公式サイト，あるいは熱狂的なファンやファンに一元的に情報を提供するページが存在するであろう．このとき，各公式サイトはファンのページから多くのリンクを集め，逆にファンのページはそのような公式サイトに数多くのリンクを張っていることが想像される．そしてこのオーソリティとハブは，互いにその性質を高めあう相互強化関係にあると考えた．

HITS は具体的には，ウェブグラフ $G$ 全体に対してではなく，クエリの検索キーワード $q$ によって限定された**注目部分グラフ** $G_q$ に対して動作する．まず始めに，テキストベースの検索エンジンによる検索結果を利用して，検索キーワードに合致するページの中で関連性が高いページを（200 ページ程度）集めて種ページ集合とする．これをもとに，この種ページ集合に属するページから直接あるいは数ステップ以内のリンクをもつページをたどり一定数のページを集める．これがこのキーワードに関連するトピックの権威的なページのほぼすべてを含むとみなし，これらのページが誘導する部分グラフを $G_q$ とする（したがって HITS はこの時点でクエリ従属である）．

次にこの $G_q$ に属する各ページ $v$ に対して，オーソリティスコア $x_v$ とハブスコア $y_v$ の2種類のスコアを計算し，オーソリティベクトル $\boldsymbol{x}=(x_1,\ldots,x_k)^\top$ とハブベクトル $\boldsymbol{y}=(y_1,\ldots,y_k)^\top$ を求める $(k=|V(G_q)|)$．それらは，各ページ $v$ の初期ハブスコア $y_v^{(0)}$ を与えたうえで，以下の式で反復計算される：

$$\begin{cases}\overline{x}_v^{(i+1)}=\displaystyle\sum_{u\in N^-(v)} y_u^{(i)},\ \boldsymbol{x}^{(i+1)}=\overline{\boldsymbol{x}}^{(i+1)}/\|\overline{\boldsymbol{x}}^{(i+1)}\|_2,\\ \overline{y}_v^{(i+1)}=\displaystyle\sum_{w\in N^+(v)} x_w^{(i+1)},\ \boldsymbol{y}^{(i+1)}=\overline{\boldsymbol{y}}^{(i+1)}/\|\overline{\boldsymbol{y}}^{(i+1)}\|_2.\end{cases}\tag{1}$$

ただし $\|\cdot\|_2$ は $L_2$ ノルムを表し，各回の反復でのスコアはその2乗和が1になるように正規化されている．

この式（1）で（正規化前の）各回のオーソリティスコア $\overline{x}_v$ は，そのページ $v$ へリンクを張るページ $u$ のハブスコアの和として，ハブスコア $\overline{y}_v$ は，そのページ $v$ がリンクを張るページ $w$ のオーソリティスコアの和として定義されており，これが相互強化関係を表している．

この計算は，注目部分グラフ $G_q$ の隣接行列を $A$ とすると，それぞれ

$$\boldsymbol{x}^{(i+1)}=A^\top\boldsymbol{y}^{(i)},\ \boldsymbol{y}^{(i)}=A\,\boldsymbol{x}^{(i)}$$

と表される．これらは

$$\boldsymbol{x}^{(i+1)}=A^\top A\,\boldsymbol{x}^{(i)},\ \boldsymbol{y}^{(i+1)}=A\,A^\top\boldsymbol{y}^{(i)}\tag{2}$$

と変形でき，それぞれ $A^\top A$, $A\,A^\top$ の主固有ベクトルを求めるべき乗法（power method）[9][10][19] に他ならない．したがって，オーソリティベクトル $\boldsymbol{x}$ とハブベクトル $\boldsymbol{y}$ は，それぞれ $A^\top A$ と $A\,A^\top$ の主固有ベクトル（方向の単位ベクトル）を求めることに相当する．行列 $A^\top A$ と $A\,A^\top$ の主固有値 $\lambda_1$ と第2固有値（絶対値が2番目に大きい固有値）$\lambda_2$ には $|\lambda_1|>|\lambda_2|$ の関係が成り立ち，適当な初期ハブベクトル $\boldsymbol{y}^{(0)}$ $(\neq\mathbf{0})$（たとえば $\boldsymbol{y}^{(0)}=(1,\ldots,1)^\top$）から開始した式（2）の計算は，$A^\top A$ と $A\,A^\top$ の主固有ベクトルに収束することが知られている．

図1は，5頂点からなるグラフを注目部分グラフとし，各頂点のオーソリティスコアとハブスコアを示している．オーソリティスコアに注目すると，A や D, E が高い値をもつ．ハブスコアにも注目すると，A, D をオーソリティ集合，C, E をハブ集合とする部分グラフ（太矢印）が，潜在する密な2部グラフ（の一つ）とし

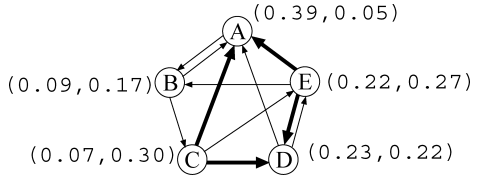


**図1** 小さな注目部分グラフと各ページ $v$ のオーソリティスコア，ハブスコアの組 $(x_v,y_v)$

  

てとらえられることがわかる．

HITS は，2003 年に teoma という検索エンジンがそのアイデアを採用し，さらに teoma を採用した ask.com という検索サイトが実用化した．日本市場にも ask.jp として鳴り物入りで参入し，個人的にも大きな期待をもって試用したが，そのランキングの網羅性や的確さには大きな問題があったと言わざるを得ない．その原因はランキング手法が部分的にクエリ従属であること，ひいてはハブ-オーソリティ・モデルの妥当性にあると考える．結局，ask.jp は 2009 年に日本市場から撤退した．

その一方で，ハブ-オーソリティ・モデルの仮説を認めれば，逆にウェブグラフ中に潜在する密な 2 部グラフ構造を見いだすことで，ウェブ上で，あるトピックに興味をもつコミュニティを発見できるのではないかと考えられる [16]．このように，ウェブを巨大なデータベースとみなして隠れた情報を見つけることをウェブマイニングと呼び，Kleinberg の成果はその後のウェブマイニングに対するリンク解析のアプローチによる研究の先駆となった [13][29]．

### 4.2 PageRank

Page と Brin が考案した PageRank [27] は，次のような単純なアイデアに基づく：ランクが高いページとは，より多くのしかもランクの高いページからの厳選されたリンクを受けているページである．このアイデアを表現するために彼らが導入した式は，ページ $v$ のページランクスコア $r_v$ を，

$$r_v = \sum_{u \in N^-(v)} \frac{r_u}{|N^+(u)|} \quad (v \in V)$$

という線形等式系の，$\sum_{v \in V} r_v = 1$ という正規化の制約のもとでの解とするものである．しかしこのままの定義にはいくつかの不都合があり，それらを解消するために彼らが施した変形は，いまでは以下のように説明される．

ウェブグラフの $n$ 点を状態とし，グラフの隣接行列をもとに定義される行列 $\mathbf{G}$ を推移確率行列とするマルコフ連鎖を考え，ページのランクをその定常分布とみなす．具体的には，$A = (a_{uv})$ をウェブグラフの隣接行列とするとき，これを既約な推移確率行列とするために，

$$p_{uv} = \begin{cases} a_{uv}/|N^+(u)| & (N^+(u) \neq \emptyset), \\ 1/n & (N^+(u) = \emptyset), \end{cases}$$

で $P = (p_{uv})$ を定義する．すなわち，隣接行列において非零成分が存在する行は行和が 1 となるように正規化する．またそうでない行はリンク先のないページ（dangling page）に対応するので，これを解消するためすべての要素を $1/n$ とし，マルコフ連鎖を既約にする．そのうえで，全成分が 1 である $n$ 次元ベクトル $\boldsymbol{e}$（ここでは単位ベクトルではないので注意）を用いて行列 $\boldsymbol{e}\boldsymbol{e}^\top$（すなわち全成分が 1 の $n \times n$ 行列）を準備し，$\mathbf{G}$ を $P$ と $n^{-1}\boldsymbol{e}\boldsymbol{e}^\top$ の凸結合

$$\mathbf{G} = \alpha P + (1-\alpha)\, n^{-1}\, \boldsymbol{e}\, \boldsymbol{e}^\top \tag{3}$$

$(0 < \alpha \leq 1)$ として定義する．（$\mathbf{G} = (g_{uv})$ の各成分は，もとの隣接行列 $A$ を用いて $P$ を経ずに直接 $g_{uv} = \alpha |N^+(u)|^{-1} a_{uv} + (1-\alpha) n^{-1}$ と書くこともできる．）

ここで $\alpha$ は減衰率（damping factor）と呼ばれる係数で，式（3）の右辺は，閲覧者は確率 $\alpha$ でページ内のリンクを等確率でたどり，確率 $1-\alpha$ でリンクとは無関係に任意のページに等確率でジャンプすることを表している．これは，ウェブページの閲覧者（サーファー）の典型的な閲覧行動をモデル化しているという直観的な解釈ができ，**ランダムサーファー・モデル**と呼ばれる．また，式（3）の右辺第 2 項中の $n^{-1}\boldsymbol{e}^\top$ は**テレポーテーションベクトル**と呼ばれる．このモデルでは，結果的にページにおける滞在時間がページの重要度を表していることになる．減衰率 $\alpha$ を変化させることで異なるランクが得られるが，Page と Brin が提案した $\alpha = 0.85$ が良好な結果を与えるとされる．

行列 $\mathbf{G}$ は原始的であり，$\mathbf{G}$ を推移確率行列とするマルコフ連鎖の定常分布であるページランクベクトル $\boldsymbol{r}_\alpha$ は，したがって $\mathbf{G}^\top$ の主固有ベクトルを計算することで求められる．例えば，図 2 のウェブグラフに対する各ページのページランクスコアは図中に示すとおりとなる．ただし，減衰率 $\alpha = 0.85$ としている．多くのリンクを集めるページ A と，A からの唯一のリンクを受けるページ B のランクが高いことがわかる．また，ページランクで最高のスコア 0.32 を得るページ B

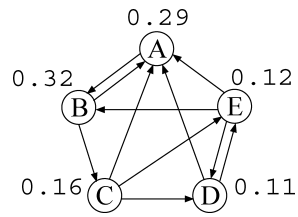


**図 2** 小さなウェブグラフと各ページのページランクスコア

の，HITS でのオーソリティスコアは最低に近い 0.09 であり，PageRank と HITS のランキングの性質が大きく異なることが確認できる．

ページランク（行列 $\mathbf{G}^{\mathrm{T}}$ の主固有ベクトル）の計算は，実際にはこれもべき乗法で行われており，実用的には 50〜100 回程度の反復で収束すると言われている．べき乗法の漸近的な収束の速さは，$\lambda_1, \lambda_2$ をそれぞれ主固有値，第 2 固有値として $|\lambda_2/\lambda_1|^k \to 0$ の速さに等しく，行列 $\mathbf{G}^{\mathrm{T}}$ に対しては $|\lambda_2| \leq \alpha$ が知られている [3][17][18]．したがって，$\alpha$ が 1 に近づくにつれて収束の速度は極端に遅くなる可能性があることから，その意味でも減衰率 $\alpha = 0.85$ という選択は都合がよいと考えられる．

発明者の起業の才覚もあり，PageRank の原理が Google という商用検索エンジンとなり，IT 企業としても大成功をおさめている事実はみなが知るところである．ページランクを導く行列 $\mathbf{G}$ は，俗に**グーグル行列**と呼ばれている．もちろん，いまやページランクはページの重要度を測る一つの尺度にすぎず，これ以外にコンテンツに基づく古典的なものも含めて 200 以上もの基準で最終的な重要度を決定しており，しかもそのアルゴリズムは週単位で更新される．Google は，そのブログで 1 兆ページ以上をインデックス化したと報告している（2008 年 7 月 26 日）．また，主要なページに関する情報は数分〜数時間の頻度で更新されていることが確認できる．さらに PageRank がクエリ独立で，ランクを事前に（オフラインで）計算できることは応答速度に有利に働き，検索結果は人のまばたきの時間を目安に 0.25 秒以内に返すことを目標としている．

このように Google は，重要度だけではなく，網羅性，リアルタイム性，応答速度などいずれの項目においても進化を続ける．これらに加えて重要な事実は，Google がスパムをほぼ完全に排除することに成功していることである（これにも PageRank のクエリ独立性が有利に働く）．このことは検索結果の信頼性を大きく高め，前述の項目とも合わせて Google がキーワード検索連動型広告を収入源とするビジネスモデルを確立する原動力となった．またこのようなアイデアを実現するためには高度なハードウェア技術が必要で，Google はそれらを開発し持ち合わせている事実も見逃せない [26]．

## 5. 発展的な話題

前節で見たように，リンク構造に基づきウェブページのランクを計算する 2 つのアイデアは，数学的には行列の主固有ベクトルの計算に帰着される．しかしながら，意味のある実用をもつ行列のサイズとしては世界最大（1 兆 ×1 兆以上）であろう超大規模行列の実際の計算には，机上だけではすまないさまざまな議論をともなう [3][6][17][18]．理論的には，例えば固有ベクトルを数値的に計算する効率的な方法やその収束性，収束の速さなどが興味の対象となる．その中で，単純なべき乗法が総合的に優位であることは興味深い．実装上はストレージ，分散計算などの観点が求められる [21][26]．Google は全ウェブページのランクを計算するのに 3 日〜1 週間を費やすとされており [2][11]，検索結果のリアルタイム性のためにも，ソフトウェア，ハードウェア両面からその計算の高速化は重要である．

以下では，これら以外の事項に触れる．

### 5.1 ランクの安定性

ランキングアルゴリズムが 1 つのウェブページに与える重要度は，その内容に変更がなければ大きく変化すべきではない．しかしながら，リンク構造に基づくランキングは，ページ内容に変更がなくとも構造の変化がスコアに大きな影響を与える可能性がある．実際，ハードウェアの故障やクローラの動作が原因で一時的にリンクやページが欠損するなど，リンク構造はむしろ常に完全ではない．このような状況で，ランキングアルゴリズムが計算するランクは安定であることが求められる．

これに関して，HITS は非常に不安定であることが知られている [7]．例えば，本来のリンク構造が図 3 左である注目部分グラフと，そこから太矢印のリンクが 1 本欠損した図 3 右に対して，各頂点のオーソリティスコアを図中に示す．このとき右上の 2 頂点のスコアは大きく変化し，HITS によるスコアが非常に不安定であることがわかる．

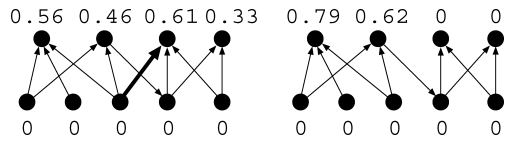


**図 3** 微小なリンク構造の違いをもつ 2 つの小さな注目部分グラフと各ページのオーソリティスコア

一方 PageRank については，リンク構造の変化などでページのスコアが変化したとしても，変化前後のページランクベクトル $\boldsymbol{r}^{\top}$ と $\tilde{\boldsymbol{r}}^{\top}$ に対して次式が成り立つことが示されている [23]:

  

$$\|\boldsymbol{r}^\top - \tilde{\boldsymbol{r}}^\top\|_1 \leq \frac{2\alpha}{1-\alpha}\sum_{v\in U} r_v.$$

ただし $U$ はランクが更新されたページ集合，$\|\cdot\|_1$ は $L_1$ ノルムを表す．スコアの変動が微小であれば全体への影響は小さくランクは安定であることがわかり，ここでも PageRank に優位性があるとともに，安定性の観点からも減衰率 $\alpha$ は 1 よりある程度小さい（$\alpha = 0.85$ のような）設定が望ましいと言える．

### 5.2 パーソナル化ランキング

同じ検索キーワードによる検索結果はすべてのユーザに対して同一である必要はなく，むしろ個人の嗜好を反映してユーザごとに異なるべきである．例えば，“apple” というキーワードによる検索は，リンゴ栽培農家か，コンピュータ製品の購買予定者か，万有引力法則の学習者によるものかわからない．このように，検索結果をユーザごとに調整するパーソナル化ランキングは，実用的に重要な方向性の 1 つであると考えられている [3][8][12][17].

PageRank に個人の嗜好を反映することは原理的には単純で，そのアイデアは当初から Page と Brin [27] により提案されていた．それは，グーグル行列 $\mathbf{G}$ におけるテレポーテーションベクトル $n^{-1}\,\boldsymbol{e}^\top$ を，必ずしも一様ではない確率ベクトル $\boldsymbol{v}^\top$ に変更するというもので，これを**パーソナル化ベクトル**と呼ぶ．ここでは，例えばよく訪れるウェブページへの推移確率を大きくする．しかしながら，ウェブ全体に対する計算に数日かかるページランクを，ユーザごとに計算し誂えることは計算負荷の観点から実現が困難である．

ところがこの発想は，リンクに基づくランキングに対するスパムを除去する方法を与えるという副産物を生んだ．ページ外要素であるリンクによるランキングのスパムは，例えば次のようにして発生しうる．図 2 で最低のページランクスコアを得たページ D がそのスコアを上げようとするとき，D を含む密な構造（ここでは，4 頂点からなる双方向クリーク）を作る（図 4）．するとそのスコアは 0.11 から 0.156 へと上昇し，新しく加えられた三つのページも 0.104 のスコアを得ることで，ページ A, B, C, E のスコアは相対的に大きく減少する．これはもし D を含むサイトの管理者であれば意図的に可能であり，そうでなくともブログのような SNS などでは自然発生しやすいとされる構造が高いランクを受けることになり，問題視されている．Google では，信頼できるサイトへの推移確率を人為的に上げる変更をグーグル行列のパーソナル化ベクトルに加えることでこのようなページに不当に高いランクを与えないようにし，スパムに対応しているようである．Google はこの技術を TrustRank と呼んでいる．

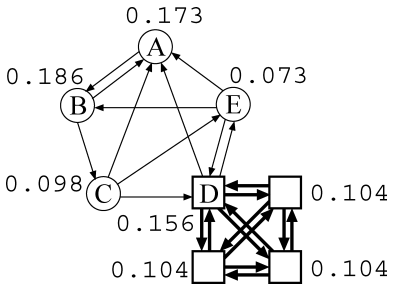


**図 4** リンクによるランキングのスパムと各ページのページランクスコア

パーソナル化ランキングを実際に実現する方法としては，個人の検索履歴やブックマークを利用し，過去に訪れたページやそれに関連するページを調べ，それらのランクを応答時に直接上げるのが実情のようである．また Google News サービスでは，興味があるキーワードを自分で追加することで，関連トピックに関するニュースをヘッドラインに表示させる仕組みも実現されている．このような検索のパーソナル化は，個人情報保護に関する問題をはらむ．

### 5.3 展望

検索結果のパーソナル化は，理論的にも応用的にも発展の余地が大きい [28]．とくに，地理情報と連動したパーソナル化はローカルサーチと呼ばれ，今後も高い需要が見込まれる．また検索対象となるデータの種類はテキストだけではなく，音声や画像，地図，メール，書籍など多岐にわたり，今後も増加する．そこでは，たとえばアイドル名で検索すると動画が上位にランクされるような検索対象を横断したランキングも必要となる．

検索結果にはキーワードに連動して広告が表示されることが多いが，この広告枠に競争原理が働くと，より大きな広告収入が発生する．近年 Google は，広告枠の割当てにオークションを導入しており，その理論を研究するメカニズム・デザインの実験的で重要な実例として注目される [24][25]．検索サイトとしては，興味をもつトピックに関連する良質の広告を検索結果に付随して得ることを目的に，ユーザが検索を利用するのが理想だという．

## 6. おわりに

リンク構造に基づくウェブページのランキング技術を概観したが，個人的には現在の Google の検索結果

に不満を感じることはほとんどなく，その技術は完全で，とくにそれを支えるPageRankの原理は絶対であるように見える．別の見方として，ウェブページやリンクは自然発生的で，ある種の自然現象である．自然現象を説明する法則は単純であるべきである．ページランクを与えるグーグル行列の定義は，ウェブの構造がもつランダムネスを自然に表現したうえに，単純で美しい．その美しさにはランキングの真理があるように見えるのだが，将来これに優るランキング手法は出現するだろうか．一方，検索はあくまでも手段にすぎず，的確な検索結果を得るための検索キーワードを入力するのも人間なので，最終的には人間の智慧や知識が重要であることも明白である．

## 参考文献


[1] Y. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke and S. Raghavan, "Searching the web," *ACM Transactions on Internet Technology*, Vol. 1, pp. 2–43 (2001).

[2] アスキー．進化する検索エンジン．月刊アスキー，2005年5月号 (2005).

[3] P. Berkhin, "A survey on PageRank Computing," *Internet Mathematics*, Vol. 2, pp. 73–120 (2005).

[4] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, Vol. 33, pp. 107–117 (1998).

[5] A. Z. Broder, S. R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. L. Wiener, "Graph structure in the web," *Computer Networks*, Vol. 33, pp. 309–320 (2000).

[6] G. M. Del Corso, A. Gullí and F. Romani, "Fast PageRank computation via a sparse linear system," *Internet Mathematics*, Vol. 2, pp. 251–273 (2006).

[7] D. Fogaras, "Algorithms on the web graph," *Proc. 3rd HJ Symposium on Discrete Mathematics and its Applications*, pp. 240–249 (2003).

[8] D. Fogaras and B. Rácz, "Towards scaling fully personalized PageRank," *Proc. 3rd WAW*, pp. 105–117 (2004).

[9] J. B. Fraleigh and R. A. Beauregard, "*Linear Algebra*" (3rd ed.), Addison-Wesley (1995).

[10] G. Golub and C. F. van Loan, *Matrix Computations*. John Hopkins University Press (1989).

[11] Google株式会社（監修）．Googleサービス徹底解剖．インプレス (2006).

[12] T. Haveliwala, "Topic-sensitive PageRank," *Proc. 11th WWW Conf.*, pp. 517–226 (2002).

[13] M. R. Henzinger, "Algorithmic challenges in web search engines," *Internet Mathematics*, Vol. 1, pp. 115–126 (2003).

[14] J. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, Vol. 46, pp. 604–632 (1999).

[15] J. Kleinberg and S. Lawrence, "The structure of the Web," *Science*, Vol. 294, pp. 1894–1895 (2001).

[16] R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, "Trawling the Web for emerging cyber-communities," *Proc. 8th WWW Conf.*, pp. 403–416 (1999).

[17] A. N. Langville and C. D. Meyer, "Deeper inside PageRank," *Internet Mathematics*, Vol. 1, pp. 335–380 (2005).

[18] A. N. Langville and C. D. Meyer, *Google's PageRank and Beyond*, Princeton University Press (2006). (岩野和生，黒川利明，黒川洋 訳，PageRankの数理．共立出版 (2009).)

[19] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM (2000).

[20] M. Mitzenmacher, Editorial: The future of power law research, *Internet Mathematics*, Vol. 2, pp. 525–534 (2006).

[21] 村田剛志（編）．検索エンジン2005―Webの道しるべ―．情報処理，Vol. 46 (2005).

[22] M. E. J. Newman, The structure and function of complex networks, *SIAM Review*, Vol. 45, pp. 167–256 (2003).

[23] A. Ng, A. Zheng and M. Jordan, "Link analysis, eigenvectors and stability," *Proc. 7th IJCAI Conf.*, pp. 903–910 (2001).

[24] N. Nisan, "Introduction to Mechanism Design (for Computer Scientists)," In *Algorithmic Game Theory*, Cambridge (2007).

[25] N. Nisan, "Google's auction for TV ads," *Proc. 36th ICALP, Part II*, pp. 309–327 (2009).

[26] 西田圭介．Googleを支える技術．技術評論社 (2008).

[27] L. Page, S. Brin, R. Motwani and T. Winograd, "The PageRank citation ranking: Bring order to the Web," *Technical Report of the Stanford Digital Library Technologies Project* (1998).

[28] D. Sontag, K. Collins-Thompson, P. N. Bennett, R. W. White, S. Dumais and B. Billerbeck, "Probabilistic models for personalized web search," *Proc. 5th ACM WSDM Conf.*, pp. 433–442 (2012).

[29] 宇野裕之．ウェブグラフ―その性質と利用―．OR研究の最前線，日本オペレーションズ・リサーチ学会誌，Vol. 51, pp. 757–763 (2006).


  