ノードランキング手法の組み合わせによる人間が直感で付けたノードランキングの再現と考察

本庄将也* 鈴木育男 山本雅人 古川正志
(北大工)† (北大情報科学)‡

1 はじめに

近年，WWW は拡大し続けており，2006 年 2 月時点で静的なページだけでも 150 億ページ以上，動的に生成されるページを含めると 350 億ページ以上が存在するとの推定がある [1]．この大規模なネットワークから目的のページを見つけることは困難であり，様々なアプローチで研究されている．その中の 1 つにキーワードマッチングとページの重要度ランキングを組み合わせて検索する手法がある．この方法では，ページの重要度ランキングをどのようにするかが重要であり，利用者が目的としているページを上位にランキング付けするアルゴリズムの研究が盛んに行われている．

本研究ではその中でもよく知られたアルゴリズムである HITS[2] と PageRank[3]，近傍ノードのリンク次数を利用したノードランキング法を組み合わせることで，人間が直感で選んだノードランキングを再現できるかを検証し，人間がどのようにランキング付けしているか考察する．

2 HITS

HITS(Hyperlink-Induced Topic Search) とは Jon Kleinberg によって提唱された，web ページのハイパーリンク情報を用いて web ページの重要度を算出する手法である．

HITS は“優良な authority(権威のある web ページ) は多くの優良な hub (リンク集 web ページ) からのリンクを持っていて，優良な hub は多くの優良な authority へのリンクを持っている”という仮説に基づいている [3]．

この authority と hub の関係性から，authority と hub は数値的には相互補強関係があると考えられ，authority 値と hub 値の更新式は以下のように定義される [2]．

$$x^{\langle p \rangle} \leftarrow \sum_{q:(q,p)\in E} y^{\langle q \rangle} \tag{1}$$

$$y^{\langle p \rangle} \leftarrow \sum_{q:(p,q)\in E} x^{\langle q \rangle} \tag{2}$$

ここで，$x^{<p>}$ は web ページ p の authority 値，$y^{<p>}$ は web ページ p の hub 値，$q:(q,p)\in E$ は web ページ q から web ページ p へのリンクが存在するすべての web ページ q とする．

(1) と (2) の演算と値ベクトルの正規化を繰り返し，値ベクトルが収束したとき，それをそれぞれの web ページの authority 値と hub 値とする．

3 PageRank

PageRank とは，Google 社によって開発された，web ページのハイパーリンク情報を用いて web ページの重要度を算出する手法である．

PageRank では“優良な web ページは優良な web ページからのリンクを持っている”という考えを使い，web ページの重要度ベクトル x を次の更新式を繰り返し適用させて求めている．

$$\mathbf{x} = B\mathbf{x} \tag{3}$$

行列 B は 2 つの web ページへのアクセス方法をモデル化している．1 つ目は web ページ上のハイパーリンクを使い移動する方法で，次の式でモデル化される．

$$b'_{ij} = \frac{a_{ji}}{\sum_k a_{jk}} \tag{4}$$

ここで，a_{ij} は隣接行列 A の (i,j) 成分である．2 つ目はハイパーリンクを使わずに移動する方法 (アドレスの直接入力やブックマーク等) で，web ページ数 n を用いて，各々確率 $1/n$ で移動すると考える．これらを合わせて B は次のように定義される．

$$B = \alpha B' + (1-\alpha)\frac{1}{n}ee^T \tag{5}$$

ここで，$e=(1,1,\cdots,1)^T$ であり，$\alpha = 0.8$〜0.9 が良いとされている [3]．

(5) の演算を繰り返し，値ベクトルが収束したとき，それをそれぞれの web ページの PageRank スコアとする．

4 近傍ノードのリンク次数を利用したノードランキング法

近傍ノードのリンク次数を利用したノードランキング法とは，注目ノードとその隣接ノードの入次数，又は出次数の順にランキング付けを行う手法である．

注目ノード p とその隣接ノードの入次数 dd_{in_p} と出次数 dd_{out_p} は以下の式により定義される．

$$dd_{in_p} = d_{in_p} + \sum_i \sum_j a_{ji}a_{ip} \tag{6}$$

$$dd_{out_p} = d_{out_p} + \sum_i \sum_j a_{pi}a_{ij} \tag{7}$$

ここで，d_{in_p} はノード p の入次数，d_{out_p} はノード p の出次数，a_{ij} は隣接行列 A の (i,j) 成分である．

honjyo@complex.eng.hokudai.ac.jp
札幌市北区北 14 条西 9 丁目北海道大学工学部
札幌市北区北 14 条西 9 丁目北海道大学大学院情報科学研究科

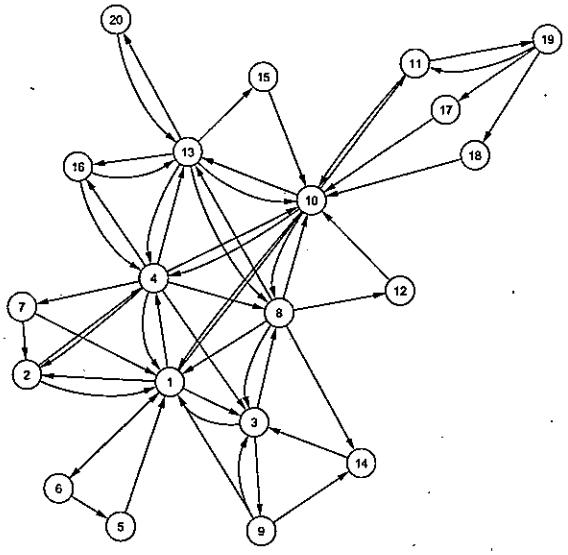

Fig. 1 実験で使用するネットワーク図

5 実験及び実験結果の考察

本実験では，ネットワークに対して HITS と PageRank，及び近傍ノードのリンク次数を利用したノードランキング法を適用し，それぞれのノードの重要度を比較する．また，それぞれの手法で得られたノードの重要度を組み合わせて，人間がそのネットワークを見て直感で選んだノードランキングを作ることができるか実験する．

5.1 各アルゴリズムの比較実験

図 1 のネットワークに対して，HITS と PageRank，及び近傍ノードのリンク次数を利用したノードランキング法を適用し，得られた重要度を比較した．

重要度は，HITS の authority 値と hub 値，PageRank の PageRank スコアと隣接行列 A を転置してから PageRank を適用したときの PageRank スコア，近傍ノードのリンク次数を利用したノードランキング法の dd_{in} と dd_{out} の 6 つを用いた．なお，これらの値は以下では HITS(auth)，HITS(hub)，PageRank(A)，PageRank(A^T)，dd_{in}，dd_{out} と表記する．

5.1.1 実験条件

実験条件は以下のように設定した．

検証するネットワーク

ノード数 20，エッジ数 56 のネットワーク (図 1) を用いる．

アルゴリズムのパラメータ

PageRank のパラメータ α は $\alpha = 0.85$ とした．

HITS と PageRank の固有値計算

固有値計算には冪乗法を用い，反復計算の中で計算前後の固有ベクトルの値が有効数字 3 桁が一致するときに終了した．

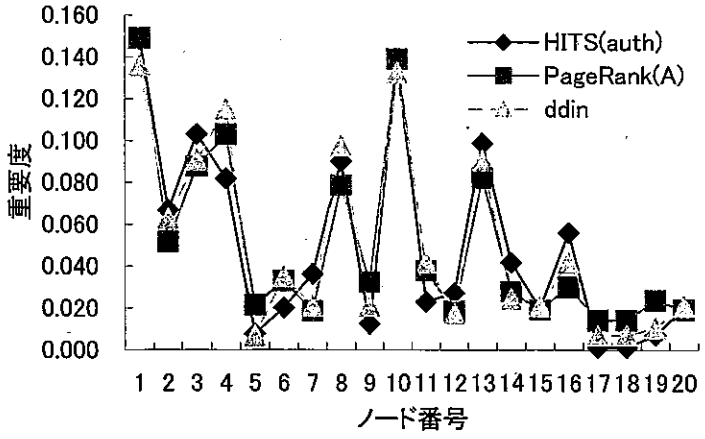

Fig. 2 HITS(auth) と PageRank(A) と dd_{in} の重要度

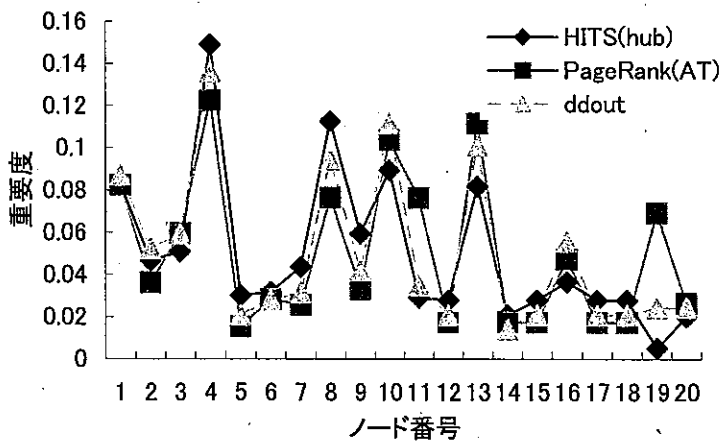

Fig. 3 HITS(hub) と PageRank(A^T) と dd_{out} の重要度

ノードの重要度

各アルゴリズムでノードの重要度の値の幅が異なるので，ノード i の重要度 v_i が $0 \leq v_i \leq 1$ かつ $\sum_i v_i = 1$ となるように正規化した値を用いる．

5.1.2 実験結果

図 2，及び図 3 に実験結果を示す．

5.1.3 考察

図 2 と図 3 を見るとそれぞれグラフの概形は一致しており，図 2 は対象ノードへの注目度によるランキング，図 3 は対象ノードからの遷移自由度によるランキングと分類することができる．

一方，細部を見ると，図 2 はノード 4，9，16 などで，図 3 はノード 8,9,19 などで大きな値の差が見られる．これは，HITS は hub 値の高いノードからリンクをされていれば authority 値が高くなり，PageRank は PageRank スコアが高いノードからリンクされていれば PageRank スコアが高くなり，dd_{in}，dd_{out} は注目ノードとリンクし

ID	Point							Importance	Rank
	a	b	c	d	e	f	average		
1	7	8	8	8	8	7	7.667	0.213	1
2	3	2	2	2	2	3	2.333	0.065	7
3	2	4	3	3	5	0	2.833	0.079	6
4	4	6	7	5	6	5	5.500	0.153	3
5	0	0	0	0	0	0	0.000	0.000	12
6	0	0	0	0	0	0	0.000	0.000	12
7	0	0	0	0	0	0	0.000	0.000	12
8	6	3	4	4	4	4	4.167	0.116	5
9	0	0	1	0	0	0	0.167	0.005	9
10	8	7	6	7	7	8	7.167	0.199	2
11	1	0	0	0	0	0	0.167	0.005	9
12	0	0	0	0	0	0	0.000	0.000	12
13	5	5	5	6	3	6	5.000	0.139	4
14	0	0	0	0	0	1	0.167	0.005	9
15	0	0	0	0	0	0	0.000	0.000	12
16	0	1	0	1	1	2	0.833	0.023	8
17	0	0	0	0	0	0	0.000	0.000	12
18	0	0	0	0	0	0	0.000	0.000	12
19	0	0	0	0	0	0	0.000	0.000	12
20	0	0	0	0	0	0	0.000	0.000	12

Fig. 4 アンケート結果から得られるノードの重要度

ているノードのリンクの広がりに影響される特徴があるためと考えられる.

5.2 アンケート結果と一致する組み合わせノード重要度の調査実験

図1に対して，人間の直感でノードの重要度を決めて得られたランキングと一致するように，5.1の実験で得られたノードの重要度を組み合わせることができるか検証する.

5.2.1 実験条件

使用する直感から得られたランキング

6人に，図1に対して，リンクを受けることの方が重要である場合，どのノードが重要であるかを1位から8位までランキング付けしてもらい，1位は8ポイント，2位は7ポイント,…，8位は1ポイント，それ以降は0ポイントとし，それらの平均値を $0 \leq v_i \leq 1$ かつ $\sum_i v_i = 1$ となるように正規化した値を人間の直感で決めたノード重要度 v_i とし，それに基づき作られたランキングを使用する.(図4) また，各手法で得られたノードランキングとアンケート結果から得られるノードランキングの比較(図5) より，人間の直感から得られたランキングと一致する手法がないことがわかる.

ノード重要度の組み合わせ方

各手法で得られたノード重要度に $0 \leq c_i \leq 1$ かつ $\sum_i c_i = 1$ となるような係数 c_i を掛け足し合わせた値を組み合わせたノード重要度とする.

その他の実験条件

その他の実験条件は5.1の実験と同じ条件にした.

Rank	ID						
	HITS (auth)	HITS (hub)	Page Rank(A)	Page Rank(A^T)	dd_{in}	dd_{out}	Result of the survey
1	1	4	1	4	1	4	1
2	10	8	10	13	10	10	10
3	3	10	4	10	4	13	4
4	13	1	3	1	8	8	13
5	8	13	13	8	3	1	8
6	4	9	8	11	13	3	3
7	2	3	2	19	2	16	2
8	16	2	11	3	11	2	16
9	14	7	6	16	16	9	9
10	7	16	9	2	6	11	11
11	12	6	16	9	14	7	14
12	11	5	14	6	9	6	5
13	6	11	19	20	15	19	6
14	15	12	5	7	20	20	7
15	20	15	15	14	7	12	12
16	9	17	20	12	12	15	15
17	5	18	12	15	19	17	17
18	19	14	7	17	5	18	18
19	17	20	17	18	17	5	19
20	18	19	18	5	18	14	20

Fig. 5 各手法で得られたノードランキングとアンケート結果から得られるノードランキング

HITS (auth)	HITS (hub)	Page Rank(A)	Page Rank(A^T)	dd_{in}	dd_{out}
0.29	0.00	0.16	0.00	0.50	0.05

Fig. 6 実験で得られた係数

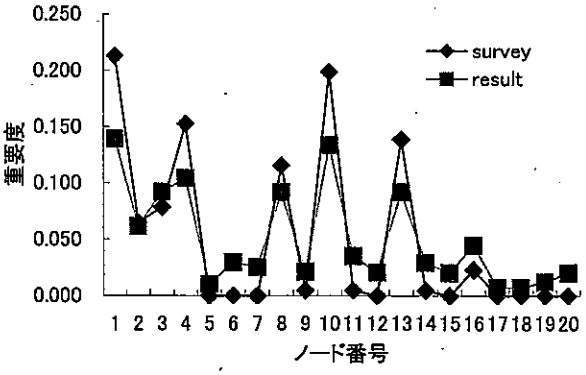

Fig. 7 アンケート結果から得られたノード重要度 (survey) と実験で得られたノード重要度 (result)

5.2.2 実験結果

図5の結果から，同じ順位のノードが存在しない1位から8位までのランキングと一致し，その中で，

$$\sum_i (importance_i - v(\mathbf{c})_i)^2 \qquad (8)$$

が最小になる $\mathbf{c} = (c_0, c_1, \cdots, c_5)$ を，条件を満たすように c_i を0.01刻みで変更して全ての組み合わせを調査した．ここで，$importance_i$ は調査結果のノードiの重要度，$v(\mathbf{c})_i$ は係数ベクトルが $\mathbf{c}$ のときのノードiの組み合わせたノード重要度である.

得られた係数ベクトル $\mathbf{c}$ を図6に示す．また，この $\mathbf{c}$ のときの $v(\mathbf{c})_i$ と $importance_i$ を図7に示す.

5.2.3 考察

以上の結果より，いくつかのノードランキング手法で得られた値を組み合わせることで，人間が直感でつけたランキングにほとんど一致させることができた．

また，得られた係数ベクトル c より，この調査条件では，人間は入次数の多いノードにそのまま注目し，その周りのノードが hub であったり，また，重要であるかどうかはあまり考えていないことがわかる．

6 おわりに

ハイパーリンクを利用したノードランキング法である HITS と PageRank，また，注目ノードとその近傍ノードのリンク次数によるランキング法を実装し，有向ネットワークに適用することで各手法の重要度の違いについて考察した．また，これらの重要度を組み合わせることで人間の直感で付けたランキングを再現できることを示し，人間がどのような基準でノードランキングを付けているか考察した．今後の課題としては，同じネットワークを異なる手法で可視化したときの人間の直感で決めたランキングの特徴の解析や，WWW などの実際のネットワークを人間がどのような基準でランキング付けしているかの研究などが挙げられる．

参考文献

[1] 加藤真, 山名早人. Fact of the Web ー 30 億ページのウェブの解析ー. DEWS2006 3B-i6, 2006.

[2] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. J. ACM, Vol. 46, No. 5, pp. 604-632, 1999

[3] Chris Ding, Xiaofeng He, Parry Husbands, Hongyuan Zha, and Horst D. Simon. PageRank, HITS and a unified framework for link analysis. In SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 353-354, 2002.