DEWS2006 2C-o3

Web ページの PageRank 値に基づくローカルコンテンツの品質推定

甲谷　優† 湯本　高行†† 小山　聡†† 田島　敬史†† 田中　克己††

† 京都大学工学部情報学科
〒 606-8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科
〒 606-8501 京都府京都市左京区吉田本町
E-mail: †{kabutoya,yumoto}@dl.kuis.kyoto-u.ac.jp, ††{oyama,tajima,ktanaka}@i.kyoto-u.ac.jp

あらまし　近年，Google [1] デスクトップ検索のように，Web 以外のローカルコンテンツを検索する機会が増えている．Google は，Web の検索結果を PageRank アルゴリズムを用いてランキングしたことで成功した．しかし，このアルゴリズムはテキストデータのようにリンク構造を持たないコンテンツを検索する際には適用できない．したがって，本研究ではそのようなリンク構造を用いたランキングアルゴリズムを適用できないコンテンツの品質を，Google の Web 検索結果から評価する手法を提案する．その結果，デスクトップ検索の結果をランキングすることができるようになり，さらに，Web とローカルという異なるリソースを横断的に検索することができるようになる．本論文では，Google の検索結果に対し本提案手法に基づいた品質評価をしてやることで，それに基づくスコアと「本来の」PageRank を比較する実験を行った．

キーワード　ランキング，ローカルコンテンツ，PageRank，類似度

Quality Estimation of Local Contents Based on PageRank Values of Web Pages

Yutaka KABUTOYA†, Takayuki YUMOTO††, Satoshi OYAMA††, Keishi TAJIMA††, and Katsumi TANAKA††

† Informatics of the faculty of Engineering, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501 Japan
†† Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501 Japan
E-mail: †{kabutoya,yumoto}@dl.kuis.kyoto-u.ac.jp, ††{oyama,tajima,ktanaka}@i.kyoto-u.ac.jp

Abstract　Recently, as Google Desktop Search, it is getting more frequent to search not web contents but local contents. Google succeeded because it used the PageRank algorithm for the ranking of the web search result. This algorithm, however, is not applicable when we search local contents without the link structure like text data etc. Therefore, in this research, we propose a method to estimate the quality of local contents, to which we cannot apply the ranking algorithm based on the hyperlink structure, by comparing them with web contents in the Google search result. As a result, it becomes possible to rank contents in desktop search result. Furthermore, this method enables to search contents across different resources such as web contents and local contents. In this paper, we applied this method to web contents in Google search results, and calculated the scores that estimate their quality. Then, we compare "estimated" PageRank score with the "real" PageRank score.

Key words　ranking, local contents, PageRank, similarity

1. はじめに

情報検索システムの検索精度を向上させるための手段として，検索結果のランキングはきわめて重要である．Web コンテンツは無数にあり，そこには有用な情報も多いが，同時に個人の日記サイトなど，人によってはまったく役に立たないコンテン

ツも多数存在する．検索システムを使うと，実行するクエリにもよるが，非常に多数の検索解が得られる．ユーザがそれら全てを１つ１つ閲覧することは考えにくく，主に検索解上位のコンテンツだけを使用するのが普通である．したがって，上位には有用な情報のみが並ぶことが望ましい．

特に，Google は重要度を自動判定する技術として「PageRank」を提案し，それを用いて検索結果をランキングすることで Web 検索サービスの分野において成功をおさめた．PageRank とは「多くの良質なページからリンクされているページは，やはり良質なページである」という概念のもと，Web ページの人気度のようなものを測定する方法である．

ほかにも，ランキングに用いられるアルゴリズムに,「HITS」と呼ばれるものが存在する．これは，多くの良質な同じ話題を持つページにリンクを貼っているページである Hub と，多くの良質な Hub からリンクされているページである Authority という概念を用いてページの品質を評価するものである．

一方で，Google デスクトップ検索のように，ローカルを対象とした検索サービスも増えてきている．かつて検索範囲が個人端末内だけで，検索対象もそれほど多くはなかったため，ローカル検索におけるランキングはさほど重要ではなかった．しかし，近年はハードディスクや DVD などの記憶媒体が急速に安価化，大容量化してきている．それにつれて，ローカルの検索範囲や検索対象も増えているため，ローカル検索における検索結果のランキングが重要になってくることが予測される．そういった Web 以外の，ローカルに蓄えられたコンテンツを，本研究ではローカルコンテンツと呼ぶことにする．

Google の成功のもととなった PageRank，Hub・Authority に基づく HITS はともに，Web のハイパーリンク構造に基づいたランキング手法である．そのため，リンク構造のないローカルコンテンツに直接適用することはできない．そこで本論文では，あるクエリが実行されたときのローカルコンテンツのランク値を，同じクエリが実行されたときの Google による Web 検索結果から，以下のように仮説を立て推定する．

内容が類似していれば，品質も近い

今回は Google における検索結果から品質を推定するため，得られたローカルコンテンツのランク値を仮想 PageRank と呼ぶことにする．

本論文は次のような構成である．まず 2 章では関連する研究と本研究との差異を述べる．3 章では仮想 PageRank を計算するのに必要となる仮説を立て，それにもとづいた計算方法を述べる．4 章では評価実験を行い，提案手法がどれほど本来の PageRank の値を反映しているかを計算する．そして正規化のためのデータを示す．5 章では，実験結果により正規化した計算式を示し，それによる仮想 PageRank の分析実験を行う．その結果から考察を行う．最後に 6 章では，本論文のまとめを行い，さらに今後の研究課題について述べる．

2. 関 連 研 究

PageRank [2] とは,「多くの良質なページからリンクされているページはやはり良質である」という再帰的な考えのもと考案された Web ページの人気度の自動測定手法である．具体的な計算方法については, [4] に詳しく書かれてある．この手法により求められたスコアが Google でのランキングに重大な影響を与えていると考えられる．本研究では，リンクをもたないコンテンツに対して，予め計算された Web コンテンツの PageRank のスコアから仮想 PageRank を計算する手法を提案する．

PageRank に関連する研究として，Cho らによる Page Quality の研究がある [5]．これは，できたばかりのコンテンツは PageRank の性質上どんなに良質でも上位には来ることができないことを問題として提起したものであり，リンク構造を持たないコンテンツに着目した我々の研究とは異なる．PageRank 値の単時間あたりの変動から，そのページの質がわかるというものである．

横断的に検索する際，本来の PageRank と仮想 PageRank の値を比較しなければならない．そのため，両者の基準を一致させるために正規化する必要がある．Montague らは，そのような異なるランキングスコアの統合の手法として，以下に示す 3 つを挙げている [6].

（1） **Standard**
すべての検索システムで出力されたスコアの最大値，最小値が等しければ，スコアは相互に比較可能であると仮定し，スコアの最小値を 0，最大値を 1 に揃える手法.

（2） **Sum**
すべての検索システムで出力されたスコアの総和が等しければ，スコアは相互に比較可能であると仮定し，スコアの最小値を 0，総和を 1 に揃える手法.

（3） **ZMUV**
すべての検索システムで出力されたスコアの平均，分散が等しければ，スコアは相互に比較可能であると仮定し，スコアの平均を 0，分散を 1 に揃える手法.

筆者は，本来の PageRank と仮想 PageRank の値の比較の際に，**Standard** の仮定に基づき，スコアの最小値と最大値を等しくする正規化を行っている．

3. 仮想 PageRank の計算手法

既存の Web コンテンツのランク値から，ローカルコンテンツのランク値をどのようにして求めるかが，本研究の研究課題である．ローカルコンテンツはリンク構造をもたないため，PageRank や HITS 等のリンク構造依存のアルゴリズムで直接スコアリングすることはできない．そこで本章では，Web 上のコンテンツのある検索システム上でのランク値からローカルのコンテンツのランク値を推測する手法について述べる．

3.1 仮想 PageRank 計算のための直観的な考え方

似たスコアを持つコンテンツ間にどのような関係が成り立つかについて，以下のような仮説を立てた．

内容が類似しているコンテンツほど，スコアが近い

この仮説を設定した理由を，以下に示す.
この仮説は，PageRank が考えられるようになった動機に基

づいている．本来，PageRank とは，日記サイトなどの個人文書のような，特定の人にしか価値のないようなコンテンツを下位に落とす目的で考えられた．このことから，たとえば Google でニュースの話題に上がったある事件について検索すると，上位にはニュースサイトのページが並び，下位にはそのニュースに関する Blog 記事などが並ぶと考えられる．その事から，似たコンテンツはランクが近いということが推測できる．

3.2 仮説に基づく仮想 PageRank の計算手法

本節では，先ほど述べた仮説に基づく仮想 PageRank の具体的な計算手法について述べる．

3.2.1 文書の特徴ベクトル化

文書の特徴ベクトル化には，tf/idf [7] 法を用いる．その際に文書の形態素解析を行うが，それには Sen [8] を用いた．文書を $D_i(i = 1, 2, \ldots, k)$ とする．文書 $D_1, \ldots, D_k$ 中に出現する単語を $t_j(j = 1, 2, \ldots, l)$ とする．

このとき，文書 D_i 中の単語 t_j の出現回数を $tf(D_i, t_j)$，単語 t_j の出現する文書数を $df(t_j)$ とする．その逆の inverse document frequency を

$$idf(t_j) = \log \frac{k}{df(t_j)} \tag{1}$$

とする．このとき，文書 D_i 中の単語 t_j の重みを以下の式で定義する．

$$w_{ij} = tf(D_i, t_j) \times idf(t_j) \tag{2}$$

このとき，文書 D_i の特徴ベクトルを以下のように定義する．

$$\boldsymbol{v}_i = (w_{i1}, w_{i2}, \ldots, w_{il}) \tag{3}$$

3.2.2 文書間類似度

文書間の類似度によってランク値は決まる．したがって文書間類似度の定義はきわめて重要である．本研究では，以下に示すように 2 つの類似度を定義し，それぞれを用いた場合の仮想 PageRank の精度を比較する．

1 つめの文書間の類似度を，文書の特徴ベクトル間の cosine similarity で定義する．この類似度は，文書の量は考慮しない．どんな長文でも短文でも，出現単語の偏りによって類似度が定まる．

この場合，文書 D_i と D_j の類似度は以下のようにして定義できる．ただし，$|\boldsymbol{v}|$ は $\boldsymbol{v}$ のユークリッドノルムである．

$$sim_1(D_i, D_j) = \frac{(\boldsymbol{v}_i, \boldsymbol{v}_j)}{|\boldsymbol{v}_i||\boldsymbol{v}_j|} \tag{4}$$

このとき，任意の i, j で $w_{ij} \geqq 0$ となるので，

$$\forall i,\ \forall j,\ 0 \leqq sim_1(D_i, D_j) < 1 \tag{5}$$

2 つめの類似度を，以下のように，文書間距離の 2 乗の逆数で定義する．

$$sim_2(D_i, D_j) = \frac{1}{|\boldsymbol{v}_i - \boldsymbol{v}_j|^2 + 1} \tag{6}$$

分母に 1 を足しているのは，文章のない文書に対応するためである．

この場合，文書の量も考慮されるため，構造の似た文書ほど類似していることになる．sim_1 と同様に，

$$\forall i,\ \forall j,\ 0 < sim_2(D_i, D_j) \leqq 1 \tag{7}$$

となることがわかる．

3.2.3 仮想 PageRank の計算

Google に対して，あるクエリ Q が実行されたとする．そのときの上位 N 件の検索結果を $W_i(i = 1, 2, \ldots N)$ とする．また，同じクエリ Q を実行してローカルを検索する際，検索結果となるコンテンツを $L_i(i = 1, 2, \ldots M)$ とする．

クエリ Q が実行されたときのコンテンツ C の Google における PageRank のスコアを $PR(Q, C)$ とする．また，本手法で求める仮想 PageRank の値を $PR'(Q, C)$ とする．このとき，$PR(Q, W_i)(i = 1, 2, \ldots, N)$ が本来の PageRank の値であり，$PR'(Q, L_i)(i = 1, 2, \ldots, M)$ が求めるべき仮想 PageRank となる．

先ほど述べた仮説に基づく仮想 PageRank を，$PR(Q, W_i)(i = 1, 2, \ldots, N)$ と，先節で述べた類似度を用いて定義する．

今，$W_i(i = 1, 2, \ldots, N)$ に対する PR の最小値を min，最大値を max とする．

まず，2 つの類似度を用いた実験式をそれぞれ以下のように定義する．

$$PR'_1(Q, L_j) = \frac{\sum_i sim_1(W_i, L_j) \cdot PR(Q, W_i)}{\sum_i sim_1(W_i, L_j)} \tag{8}$$

$$PR'_2(Q, L_j) = \frac{\sum_i sim_2(W_i, L_j) \cdot PR(Q, W_i)}{\sum_i sim_2(W_i, L_j)} \tag{9}$$

本研究では Web とローカルを横断的に検索する際，ランキングのためのスコアを，Web ページに対しては本来の PageRank，ローカルのコンテンツには仮想 PageRank を用いることを考えている．したがって，仮想 PageRank を PageRank に即して正規化する必要がある．その正規化については，次項で述べることとする．

式 (8)，(9) それぞれを Google の検索結果上位 N 件である $W_i(i = 1, 2, \ldots, N)$ に適用し，$PR'_k(Q, W_i)$ を求める．具体的には，

$$PR'_1(Q, W_i) = \frac{\sum_{j \neq i} sim_1(W_j, W_i) \cdot PR(Q, W_j)}{\sum_{j \neq i} sim_1(W_j, W_i)} \tag{10}$$

$$PR'_2(Q, W_i) = \frac{\sum_{j \neq i} sim_2(W_j, W_i) \cdot PR(Q, W_j)}{\sum_{j \neq i} sim_2(W_j, W_i)} \tag{11}$$

のように計算する．ただし，式 (10) において $|\boldsymbol{v_i}| = 0$ などの場合，$i = 1, 2, \ldots, N$ に対して $sim_1(W_j, W_i) = 0$ となってしまうため，そのときは

$$PR'_1(Q, W_i) = \overline{PR} \tag{12}$$

とする．ただし，$\overline{PR}$ は式 (25) に示している．

それぞれを用いて評価実験を行い，2 種類の類似度に基づ

く仮想 PageRank を比較する．ただし，仮説が成立していると仮定するならば，PR'_1 と PR'_2 の値はともに PR と正の相関関係があるはずである．ここで，実験式の精度が高いとは，PR'_1，PR'_2 それぞれを同じ Google の検索解である Web ページ $W_i(i = 1, 2, \ldots, N)$ に適用したときに，もとの PR を正確に復元する，あるいはそれにより得られた値ともとの PR との相関が強いと定義する．

3.2.4 正 規 化

Montague らが使用した正規化手法の中の 1 つ，**Standard** では，「すべての検索システムで出力されたスコアの最小値，最大値が同じであるならば，それらのスコアは相互に比較可能である」という仮定がなされている．したがって，仮想 PageRank を PageRank と比較するには，両者の最大値と最小値が等しくなければならない．

$k = 1$ *or* 2 に対して，このとき，$W_i(i = 1, 2, \ldots, N)$ を対象としたときの PR'_k の最小値，最大値をそれぞれ min'_k，max'_k とする．

そうすると，求めるべき仮想 PageRank は PR'_k を PageRank に即して正規化したものであるので，

$$PR'(Q, C) = \frac{max - min}{max'_k - min'_k}(PR'_k(Q, C) - min'_k) + min \quad (13)$$

となる．このようにすれば，

$$\min_i(PR'(Q, W_i)) = \min_i(PR(Q, W_i)) \quad (14)$$

$$\max_i(PR'(Q, W_i)) = \max_i(PR(Q, W_i)) \quad (15)$$

が成立し，最小値と最大値という観点から仮想 PageRank と本来の PageRank の基準が統一されることになる．

4. 正規化のための予備実験

前章において，2 つの類似度に基づく仮想 PageRank を定義した．その両者を比較し，そして仮想 PageRank を正規化するために，min'_k，max'_k の値を求める．まとめると評価実験における目的は以下の 3 つである．

（1） 2 つの類似度に基づく仮想 PageRank の比較

（2） 正規化のため，min'_k，max'_k がどのようにして与えられるかを知る

4.1 検索システムとそのスコア

本評価実験において，有名な Web 検索サービスである Google を用いた．そこでクエリ Q を投げたときの上位 N 件の Web ページを評価の対象とする．このとき，上位 n 件目のページ W_n の Google の PageRank のスコアを，正確な値を得ることができないため，以下のように定義する．

$$S(Q, W_n) = N - n + 1 \quad (16)$$

4.2 実 験 方 法

次のような手順で評価実験を行う．ただし，手順 (4) は正規化のために min'_k，max'_k を求めるためのものであり，手順 (5) はどちらの類似度を用いるかを決定するためのものである．ただし $Q_i(i = 1, 2, \ldots 10)$ の内容は表 1 に示した通りである．

表 1 クエリの内容

クエリ	内容
Q_1	京都大学
Q_2	城島 マリナーズ
Q_3	Google PageRank
Q_4	京都 学習塾 小 6 女児 殺害
Q_5	Firefox
Q_6	シーズー
Q_7	北海道 スープカレー
Q_8	Java
Q_9	ニュース 23
Q_{10}	iPod

（1） 10 組の問い合わせのためのクエリ $Q_i(i = 1, 2, \ldots, 10)$ を作成する．

（2） クエリを実行して，得られた上位 100 件の Web ページを得る．

（3） 得られた Web ページの，式 (10)，(11) に基づく $PR'_k(k = 1, 2)$ の値をそれぞれ計算する．

（4） $PR'_k(k = 1, 2)$ の最小値，最大値を求める．

（5） 求めた $PR'_k(k = 1, 2)$ と，本来の PageRank のスコア PR との相関を計算する．

すなわち，Web の検索解のページをローカルのコンテンツであると仮定したときの，それぞれの PR'_k の値を計算する．それらは Web 検索の解であるから，本来の PageRank のスコア PR を持っている．したがって，その本来の PageRank と PR'_k の値の相関をとることで，式 (8)，(9) に基づくそれぞれの仮想 PageRank を比較する．また，式 (16) より

$$min = 1 \quad (17)$$

$$max = N \quad (18)$$

となり，本来の PageRank の最小値 min，最大値 max はそれぞれ N によって与えられることがわかる．ゆえに，手順 (4) で求めた最小値 min'_k・最大値 max'_k に対する N の影響を調べるため，以下のような実験を行った．

（1） 先の実験で用いたクエリのうちの 1 つを選択し，これによる問い合わせを行う．

（2） N の値を 50,60,70,80,90,120,150,200 としたときのそれぞれの $PR'_j(j = 1$ *or* $2)$ の最小値，最大値，本来の PageRank である PR との相関を計算する．

4.3 評 価 尺 度

実験式である (10)，(11) それぞれの精度の高さとは，それぞれの PR との相関の強さである．得られた相関の値が高ければ高いほど，その実験式をもとに作られた仮想 PageRank の値が本来の PageRank を反映しているということになる．

ここで，本来の PageRank のスコアの平均，分散を

$$\overline{PR} = \frac{1}{N}\sum_{i=1}^{N} PR(Q, W_i) \quad (19)$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(PR(Q, W_i) - \overline{PR})^2 \quad (20)$$

表 2　PR'_1 の平均，分散，最小値，最大値，r_1

クエリ	$\overline{PR'_1}$	σ'_1	min'_1	max'_1	r_1
Q_1	52.195	4.275	32.153	68.395	-0.025
Q_2	52.688	4.277	34.939	69.431	0.125
Q_3	49.539	10.089	36.407	69.855	0.387
Q_4	53.114	7.326	31.392	74.205	0.217
Q_5	49,646	14.445	35.239	81.430	0.677
Q_6	50.038	4.535	34.278	64.200	0.283
Q_7	49.928	3.596	39.582	59.896	0.307
Q_8	52.176	3.370	39.449	59.644	0.019
Q_9	50.226	5.448	32.313	65.428	0.112
Q_{10}	48.840	6.425	21.000	63.819	0.215

表 3　PR'_2 の平均，分散，最小値，最大値，r_2

クエリ	$\overline{PR'_2}$	σ'_2	min'_2	max'_2	r_2
Q_1	48.474	7.559	5.898	81.440	0.041
Q_2	52.315	16.522	6.906	97.895	-0.003
Q_3	55.564	9.453	25.007	83.945	0.243
Q_4	51.973	13.751	3.028	95.512	0.336
Q_5	46.892	15.406	1.931	92.996	0.454
Q_6	55.020	12.874	1.275	96.667	0.143
Q_7	49.904	12.650	4.125	97.156	0.166
Q_8	48.818	12.132	10.132	95.637	0.311
Q_9	46.271	11.170	6.303	85.680	0.221
Q_{10}	53.187	3.744	45.222	72.173	0.055

とする．PR'_k のスコアの平均，分散を

$$\overline{PR'_k} = \frac{1}{N}\sum_{i=1}^{N} PR'_k(Q, W_i) \tag{21}$$

$$\sigma'^2_k = \frac{1}{N}\sum_{i=1}^{N}(PR'_k(Q, W_i) - \overline{PR'_k})^2 \tag{22}$$

とする．このとき，本来の PageRankPR と PR'_k の共分散は

$$cov_k = \frac{1}{N}\sum_{i=1}^{N}(PR(Q, W_i) - \overline{PR})(PR'_k(Q, W_i) - \overline{PR'_k}) \tag{23}$$

と求められる．すると，これらの値から本来の PageRankPR と PR'_k との相関は

$$r_k = \frac{cov_k}{\sigma \cdot \sigma'_k} \tag{24}$$

から得られる．ただし，$k = 1, 2$ である．

4.4 実験結果

4.4.1　2 つの類似度に基づく仮想 PageRank の比較

式 (10)，(11) それぞれの場合における PR'_k の平均，分散，最小値と最大値，本来の PageRank PR との相関を表 2，表 3 に示す．

ここで重要なのは，r の値であり，この値が大きければ大きいほどその類似度に基づく仮想 PageRank が本来の PageRank の値を反映しているということになる．

PR'_1，PR'_2 ともに，本来の PageRank PR との間に弱い正の相関関係があるか或いは無相関である．

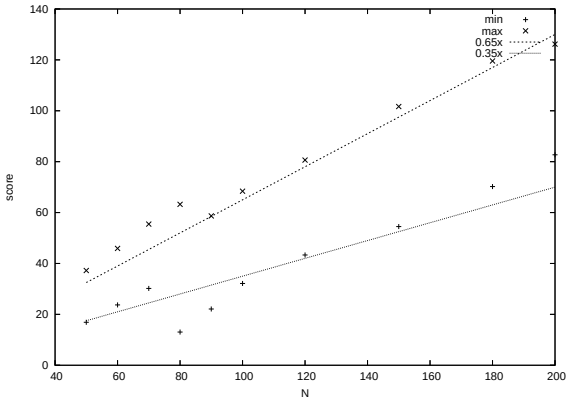

図 1　N と min'_1，max'_1 の関係

本来の PageRank のスコアの平均は，式 (25) より $N = 100$ のとき 50.5 になる．PR'_1，PR'_2 ともに，スコアの平均は 50.5 前後であり，平均値は本来のものと変わらない．

本来の PageRank のスコアの標準偏差は式 (26) より，$N = 100$ のとき 28.866 である．PR'_1，PR'_2 ともに，標準偏差は全てこの値より小さくなっている．すなわちばらつぎが小さくなっているということである．

本来の PageRank のスコアの最小値，最大値はそれぞれ 1, 100 である (式 (17)，(18))．PR'_2 の方は，最小値，最大値ともにもとの PageRank のそれと大差ない．したがって，PR'_2 は正規化せずそのまま使えるということがわかる．一方，PR'_1 の方は最小値は大きく，最大値は小さくなっている．したがって，PR'_1 を使って仮想 PageRank を求める際には正規化が必要となる．

$$\overline{PR} = \frac{N+1}{2} \tag{25}$$

$$\sigma^2 = \frac{(N+1)(N-1)}{12} \tag{26}$$

4.4.2　N と min'_1，max'_1 の関係

実験の際に用いるクエリとして，Q_1 を用いることとした．

このとき，N と，PR'_1 の最小値 $\min'_1$，最大値 max'_1 の関係を示したものが図 1 である．

この図より，min'_1，max'_1 ともに N に比例していることがわかる．すなわち，以下の等式が成り立つことがわかる．

$$min'_1 = 0.35N \tag{27}$$

$$max'_1 = 0.65N \tag{28}$$

表 2 から，いずれのクエリの場合でも上の計算式により min'_1，max'_1 が求まることがわかる．

5. 仮想 PageRank と PageRank の比較

5.1 正規化後の計算式

本来，仮想 PageRank はリンク構造を持たないローカルのコンテンツのみに適用されるべきものである．ローカルコンテンツのみを対象として検索しそれをランキングする場合は正規化の必要はなく，式 (8)，(9) によって得られるスコアをその

表 4　sim_1 に基づく仮想 PageRank

クエリ	$\overline{PR'}$	σ'	min'	max'	r'
Q_1	60.699	12.577	-1.618	109.993	-0.056
Q_2	56.619	15.783	9.890	86.715	0.331
Q_3	50.986	18.392	-4.477	94.146	0.282
Q_4	59.402	13.889	11.020	75.115	-0.032
Q_5	48.662	50.005	-10.729	178.314	0.703
Q_6	47.913	14.718	-6.811	89.960	0.284
Q_7	47.428	15.891	0.990	74.516	0.210
Q_8	57.036	10.344	23.467	77.993	-0.031
Q_9	55.893	12.264	5.580	93.884	-0.263
Q_{10}	48.707	26.822	-111.120	86.581	0.111

まま用いればよい．しかし，今後 Web とローカルという異なるリソースのコンテンツを横断的に検索したいとなると，仮想 PageRank と本来の PageRank の基準が統一されていなければならない．すなわち，W_i を対象としたときの PR' と PR の最小値，最大値が等しくなければならない．これは Montague ら [6] が提唱した **Standard** の仮定に基づいている．

5.1.1　PR'_1 をベースにした仮想 PageRank

式 (13)，(27)，(28) から，N が十分大きいとき ($N-1 \simeq N$ が成立するとき)，

$$PR'(Q,L_j) = \frac{10}{3}(PR'_1(Q,L_j) - 0.35N) + 1 \tag{29}$$

となる．

5.1.2　PR'_2 をベースにした仮想 PageRank

表 3 より，

$$min'_2 = min \tag{30}$$

$$max'_2 = max \tag{31}$$

したがって正規化する必要性はないので，PR'_2 を使って仮想 PageRank を求める場合，

$$PR'(Q,L_j) = PR'_2(Q,L_j) \tag{32}$$

となる．

5.2　仮想 PageRank の分析実験

5.2.1　仮想 PageRank と PageRank との比較実験

式 (29) を Web ページ W_i に適用する．すなわち，以下の式を用いて再びクエリ $Q_i(i = 1,2,\ldots,10)$ を用いて実験を行う．

$$PR'(Q,W_i) = \frac{10}{3}(PR'_1(Q,W_i) - 0.35N) + 1 \tag{33}$$

平均，分散，最小値，最大値，PR との相関を表 4 にまとめる．クエリが Q_1 のときはほぼ無相関で，Q_5 のときは比較的強い正の相関関係があった．クエリが Q_1，Q_5 のときの $W_i(i = 1,2,\ldots,100)$ に対する PR'_1 に基づく仮想 PageRank を求めたものが，図 2，図 3 である．

次に sim_2 に基づく仮想 PageRank だが，この結果は表 3 にある．この類似度を用いた場合はクエリが Q_5 のときにもっとも強い正の相関があり，クエリ Q_2 のときにもっとも無相関であった．

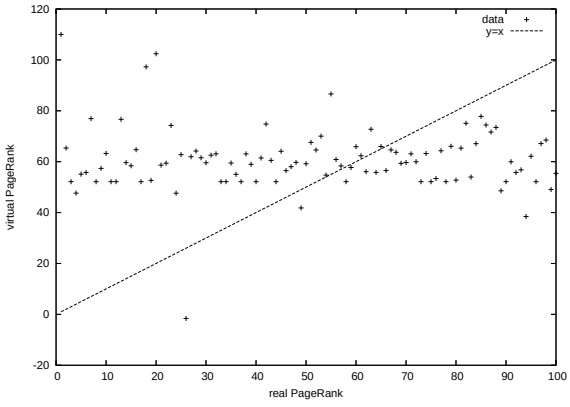

図 2　クエリが Q_1 のときの sim_1 に基づく仮想 PageRank

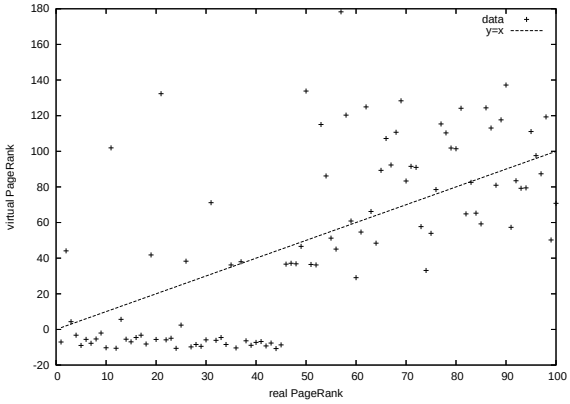

図 3　クエリが Q_5 のときの sim_1 に基づく仮想 PageRank

5.2.2　ローカル・Web の横断的検索実験

$1,2,\ldots,100$ のうち，ランダムに 30 個の異なる自然数 $a_1,a_2,\ldots,a_{30}$ を選ぶ．a 以外の自然数を $b_1,b_2,\ldots,b_{70}$ とする．このとき，$W_{a_1},W_{a_2},\ldots,W_{a_{30}}$ をローカルのコンテンツと見なし，仮想 PageRank によってランキングすることを考える．

本来の PageRank の値 PR には，式 (16) をそのまま用いる．

類似度 sim_1 を用いた場合，仮想 PageRank PR'' の計算式は，

$$PR''_1(Q,W_{a_k}) = \frac{\sum_{j=1}^{70} sim_1(W_{b_j},W_{a_k}) \cdot PR(Q,W_{b_j})}{\sum_{j=1}^{70} sim_1(W_{b_j},W_{a_k})} \tag{34}$$

を用いて

$$PR''(Q,W_{a_k}) = \frac{10}{3}(PR''_1(Q,W_{a_k}) - 35) + 1 \tag{35}$$

となる．ただし，

$$\min_j PR''(Q,W_{b_j}) = \min_j(100 - b_j + 1) \simeq 1 \tag{36}$$

$$\max_j PR''(Q,W_{b_j}) = \max_j(100 - b_j + 1) \simeq 100 \tag{37}$$

としている．

このときの $PR''(Q,W_{a_k})$ の $PR(Q,W_{a_k})$ との相関 r'' を表 5 に示す．また，クエリが Q_5，Q_9 のときの W_{a_m} に対して計算した仮想 PageRank のスコアを図 4，5 に示す．

表 5　W_{a_k} に対する PR'_1 に基づく PR'' と PR との相関 r''

クエリ	r''	クエリ	r''
Q_1	0.386	Q_6	0.339
Q_2	0.595	Q_7	0.614
Q_3	0.403	Q_8	0.040
Q_4	-0.151	Q_9	-0.108
Q_5	0.689	Q_{10}	0.337

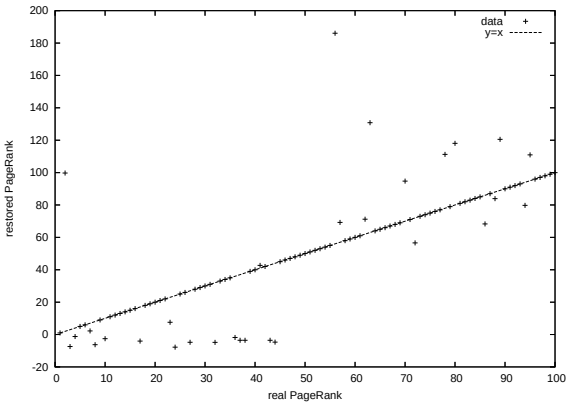

図 4　Q_5 の PageRank と sim_1 に基づく仮想 PageRank

表 6　W_{a_k} に対する PR'_2 に基づく PR'' と PR との相関 r''

クエリ	r''	クエリ	r''
Q_1	-0.400	Q_6	0.385
Q_2	0.618	Q_7	0.276
Q_3	0.282	Q_8	0.432
Q_4	-0.247	Q_9	-0.206
Q_5	0.743	Q_{10}	0.188

先項の実験とほぼ同じような結果となった．クエリが Q_5 や Q_7 のときは表 5 からは PageRank と仮想 PageRank の値がかなり近いように見えるが，実際はそうではない (図 4)．

また，類似度 sim_2 を用いた場合は，仮想 PageRank は

$$PR''(Q, W_{a_k}) = \frac{\sum_{j=1}^{70} sim_2(W_{b_j}, W_{a_k}) \cdot PR(Q, W_{b_j})}{\sum_{j=1}^{70} sim_2(W_{b_j}, W_{a_k})} \quad (38)$$

によって計算される．この場合も，式 (30)，(31) から，式 (36)，(37) が成立する．

この場合の $PR''(Q, W_{a_k})$ と $PR(Q, W_{a_k})$ の相関を，表 6 に示す．

表 6 より，クエリ Q_2 や Q_5 のときには比較的強い正の相関関係が見られる．Q_1 や Q_4，Q_9 のときのように負の相関が出る場合，両者のスコアにはほとんど関係が見られない (図 5)．

本実験において，PageRank と仮想 PageRank とで大きく評価の異なったもの，すなわち，上位のものが下位に再ランクされたもの，あるいは逆に下位のものが上位に再ランクされたものがどのようなコンテンツなのか，さらに分析を進めていく必要がある．

5.3 考　　察

仮想 PageRank と本来の PageRank の両者は異なる品質評価方法である．その中で，クエリ Q_5 などに対しては，PageRank における上位・下位の区別ができるため，仮想 PageRank が

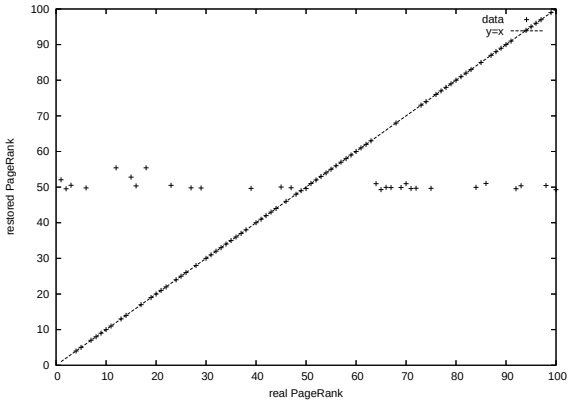

図 5　Q_1 の PageRank と sim_2 に基づく仮想 PageRank

比較的有効であったと考えられる．一方，クエリ Q_1 のような場合には，PageRank とはまったく無相関であったため，仮想 PageRank によるランキングが適切かどうかは判断できない．

5.3.1 項では，仮想 PageRank の値に大きく影響を与えると考えられるページ間の類似度の分布を分析する．また，どのように類似度が分布すれば本提案手法が有効になるのかを調べる．5.3.2 項では，2 つの類似度の性質を比較し，仮想 PageRank の計算にどちらを用いるべきか分析する．

5.3.1　PageRank と類似度の関係

仮想 PageRank のスコアは，Google 検索結果上位 100 件の Web ページ間の類似度に大きく影響を受ける．したがって，本項では Google 検索結果上位 100 件のそれぞれの ページ間類似度がどのようになっているかを分析する．

類似度の分布を表す図を示す．x 軸は左から右に Google の検索結果の上位 100 件 $W_i (i = 1, 2, \ldots, 100)$ であり，y 軸は上から下に Google の検索結果の上位 100 件 W_i を表している．色が白いほど，ページ間の類似度が高いということを表している．

クエリが Q_1 のとき，上位 100 件の Web ページ間の sim_1 は図 6 のように全体に一様に分布している．

また，本来の PageRank と比較的強い正の相関関係が見られた Q_5 のときの sim_1 では，上位間或いは下位間の類似度が高いことがわかる (図 7)．ただし，上位間・下位間においては類似度の分布は一様である．

もちろん，Q_1 のような場合でも，とってくる Google の検索結果数 N を増やせばクエリ Q_5 のときのように上位と下位のものを区別できる可能性はある．

5.3.2　2 つの類似度を用いたことにより得られた知見

出現単語の偏りのみによって決まる類似度であるコサイン相関値と，文書量も考慮した類似度として文書の特徴ベクトルの距離の 2 乗の逆数の 2 つの尺度で実験した．この 2 つで実験した理由は，PageRank でスコアが近いページとはどのように似ているかを解析するためである．

もし仮に後者の類似度を用いた仮想 PageRank の方が精度がよければ，出現単語だけでなく，文書量の近いページ，すなわち内容というよりは構造の似たページがスコアが近いという結論にいたる．例えば，ランキングの上位にはニュースサイト

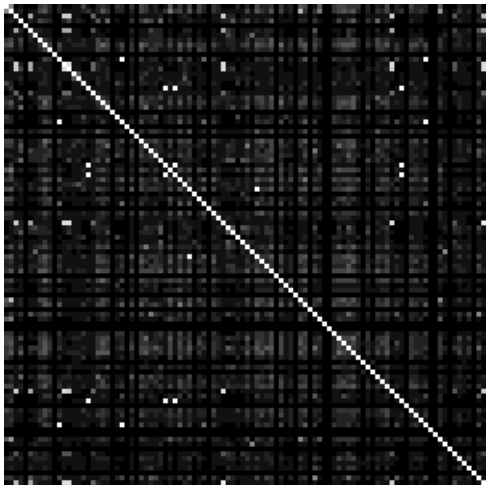

図 6　クエリが Q_1 のときの sim_1 の分布

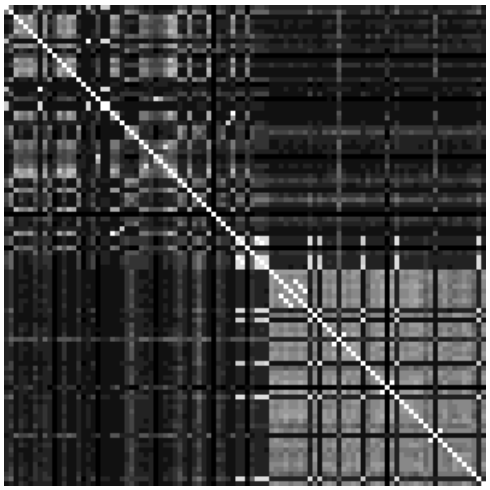

図 7　クエリが Q_5 のときの sim_1 の分布

のようなものばかりが並び，下位には Blog のような日記サイトのようなコンテンツが並ぶという結論を導くことができる．

しかし，今回の実験により文書量は PageRank と関係がないことがわかった．たとえ構造が似ていても同じようなスコアが得られるとは限らない．

このような結果が出た理由には本文のないページが影響を与えたためと考えられる．2 つめの類似度を用いると，文字列がなく画像ばかりが並んでいる 2 つのコンテンツはその内容如何に関わらず似ていると判断されてしまう．

さらに，文章量の多いページはどのページとも似ていないと判断されてしまう可能性が高い．したがって，2 つめの類似度は適切でなかったと結論できる．

6. おわりに

本論文では，類似度を用いた仮想 PageRank の計算式を考案し，それによる実験を行った．その上で実験結果から計算式を線形変換し，計算式を正規化した．

本論文では，Web ページのもつ PageRank 値をもとにリンクのないコンテンツの品質を評価する方法を提案した．その結果生まれた仮想 PageRank は，もとになっている PageRank とは異なるランキングアルゴリズムであることがわかった．

本提案手法の評価はまだできていない．今後，本提案手法に基づくランキングアルゴリズムを用いたローカルコンテンツの検索エンジンを作成し，仮説の妥当性を検証する必要がある．

今後の研究課題として，次のような点が挙げられる．

（1） 今回は，仮想 PageRank の計算に内容の類似度を用いた．しかし，本来 PageRank とは内容とは独立したコンテンツの人気度を測定する手法である．したがって，コンテンツの人気度という観点からの計算手法を考える必要がある．

（2） 今回の実験では **Standard** の仮定に基づく正規化を行った．しかし本当にこの正規化でよいのか．**SUM** や **ZMUV** の仮定に基づく正規化を試してみる価値もある．

（3） 仮想 PageRank の評価を行う必要がある．その方法としては，たとえば仮想 PageRank アルゴリズムを用いた検索エンジンを実装し，その検索精度を測るなどのものがある．

謝辞　本研究の一部は，21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」，文部科学省科学技術振興費プロジェクト「異メディア・アーカイブの横断的検索・統合ソフトウェア開発」（代表：田中克己），および，平成 17 年度科研費特定領域研究 (2)「Web の意味構造発見に基づく新しい Web 検索サービス方式に関する研究」（課題番号：16016247, 代表：田中克己）によるものです．ここに記して謝意を表すものとします．

文　献

[1] Google 日本
http://www.google.co.jp/

[2] Lawrence Page, Sergey Brin, Rajeev Montwani, Terry Winograd. *The PageRank Citation Ranking: Bringing Order to the Web*. Stanford Digital Libraries Working Paper, 1998.

[3] J. Kleinberg. Authorative sources in a hyperlinked environment. Journal of the ACM, 46(5):604–632, September 1999.

[4] Google の秘密 - PageRank 徹底解説
http://www.kusastro.kyoto-u.ac.jp/ baba/wais/pagerank.html

[5] Junghoo Cho, Sourashis Roy, Robert E. Adams: Page Quality: In Search of an Unbiased Web Ranking, *SIGMOD 2005*, pp.551-562(2005).

[6] Montague, M. and Aslam, J.: Relevance Score Normalization for Metasearch, *Proc. 10th ACM International Conference on Information and Knowledge Management* (*CIKM01*), pp.427-433(2001).

[7] Gerard Salton, *Automatic Information Organizations and Retrieval*,
McGraw-Hill(1968).

[8] Sen Project
http://ultimania.org/sen/