サイト内検索エンジンのためのスコアリング手法

伊川　洋平†　　定兼　邦彦†

† 東北大学大学院情報科学研究科　〒 980–8579 宮城県仙台市青葉区荒巻字青葉 09
E-mail: †{ikawa,sada}@dais.is.tohoku.ac.jp

あらまし　Web 検索エンジンの利便性を向上させる手段として，各ページの重要度に応じてスコアを割り当てる，Web ページのスコアリングがある．Google の PageRank は，WWW 検索エンジンで有効なスコアリング手法として広く知られているが，サイト内検索エンジンでは，PageRank のような手法ではよい結果が得られず，テキストマッチングによってのみスコアリングを行っており，Web の大きな特徴であるリンク情報を活用できていないのが現状である．そこで本論文では，Web サイトのリンク構造に特化した，サイト内検索エンジンのためのスコアリング手法である HotLink 法を発展させた HL-PR 法を提案し，有効性について検討を行う．

キーワード　Web とインターネット，情報検索

A Webpage Scoring Method for Local Web Search Engines

Yohei IKAWA† and Kunihiko SADAKANE†

† Department of System Information Sciences, Graduate School of Information Sciences, Tohoku University
Aramaki Aza Aoba 09, Aoba-ku, Sendai city, Miyagi, 980–8579 Japan
E-mail: †{ikawa,sada}@dais.is.tohoku.ac.jp

Abstract　Webpage scoring is a method to improve Web search engines that assigns a score to each page according to its importance. PageRank algorithm implemented for Google is well known as an efficient scoring method for WWW search engines, whereas it is not efficient for local Web search engines. For the latter case, text matching is usualy used for Webpage scoring, however hyperlink topology information characterizing WWW have not been well used. Although HotLink method has been proposed for scoring local Web pages, it is not well developed. In this paper, we propose HL-PR method that improves the HotLink method to give more effective ranking, and investigate it.

Key words　Web and Internet, Search Engines

1. はじめに

近年の爆発的なインターネットの普及による Web サイトの増加は，World Wide Web を巨大で有用なデータベースへと発展させた．このデータベースから効率よく情報を収集するために，多くのユーザは Google のような Web 検索エンジンを利用しているだろう．

この Web 検索エンジンの利便性を向上させる手段として，各ページの重要度に応じてスコアを割り当てる，Web ページのスコアリングがある．Web ページのスコアリングによって，ユーザは膨大なページの中からよいページを素早く探し出すことができるようになる．

Web ページのスコアリングは大別すると，ページの内容を解析し，テキストマッチングにより各キーワードに対するスコアを割り当てる手法と，Web のハイパーリンク構造を利用する手法に分類できる．本研究で扱うのは，後者のリンク構造を利用したスコアリングである．

Web のハイパーリンク構造を利用したスコアリングでは，あるページへリンクを張る行為を推薦行為とみなし，張られているリンクによってそのページの質を決定する．

自分のページのスコアが Web 全体のリンク構造によって決定するため，不正にスコアを上げることが難しく，テキストマッチングによるスコアリングと組み合わせることによって，より信頼性の高い検索エンジンを構築することができる．

リンク構造を利用したスコアリングの代表的な手法のひとつに PageRank [2] がある．PageRank は，「多くのよいページからリンクされているページは，やはりよいページである」という考え方に基づき，Web グラフのランダムウォークを単純マルコフ過程で定式化し，各ページの滞留確率をスコアとして定義する手法である．WWW 検索エンジン Google は，この PageRank を実装することによって大きな成功を収めている．

本論文は WWW 検索エンジンではなく，特定の Web サイト

内のページのみを検索することを目的とした，サイト内検索エンジンに焦点を当てている．検索したいページのある Web サイトが特定できた場合，ユーザはサイト内検索エンジンを利用することで，WWW 検索エンジンよりも確実に目的のページを検索できると期待される．

ここでは，Web サイト内のページにスコア付けを行うのに WWW 全体のリンク情報を用いずに，Web サイト内のローカルなリンク情報のみを用いて Web ページのスコアリングを行う手法について検討する．

このような条件下では，WWW 検索エンジンにおいて有効な手法として知られる PageRank は有効に働かず，サイト内検索エンジンではテキストマッチングによってのみ Web ページのスコアリングを行っており，Web の大きな特徴であるリンク情報を活用できていないのが現状である．

そこで本論文では，Web サイトのローカルなリンク構造に特化した，サイト内検索エンジンのためのスコアリング手法である HotLink 法 [4] を発展させ，HotLink 法によるスコアと PageRank によるスコアの差分をスコアとする HL-PR 法を提案し，その有効性について検討を行う．

2. サイト内検索エンジンに特有な問題点

前述のように，PageRank のような WWW 検索エンジンで有効なスコアリング手法は，サイト内検索エンジンでは有効に働かない．その原因は，両者の検索エンジンを用いて検索したい Web ページが異なるからであると考える．

ここではユーザの視点に立ち，それぞれの検索エンジンを利用して検索したいページがどのようなページなのかを議論した上で，従来の手法の問題点を明らかにしていく．

2.1 検索要求の相違

WWW 検索エンジンを利用するユーザは，無数にある Web サイトの中からキーワードと関連の深い Web サイトを検索することを目的としている．このとき，キーワードと関連の深い Web サイトに着目すると，そのサイト内のページ群の中でもっとも重要なページは，トップページであると考えられる．

たとえば「東北大学」というキーワードで WWW 検索を行うとき，多くのユーザは検索結果の上位に東北大学の Web サイトのトップページが表示されて欲しいと期待するだろう．

WWW 検索エンジンを Web サイトの検索ではなく，特定の情報が含まれている Web ページを直接検索するために利用する場合もあるが，それは専門的なキーワードや複数のキーワードを指定して，一般的なページが検索されないように検索結果を絞り込んでいるのである．

一方，サイト内検索エンジンを利用するユーザは，WWW 検索エンジンやブックマークを利用して一旦 Web サイトのトップページにたどり着いた後で，その Web サイト内のページを検索するためにサイト内検索エンジンを利用するものと考えられる．

このような状況では，ユーザにとって既知であるトップページはほとんど重要ではなく，トップページから直接リンクされているようなページも，トップページから容易に発見できるために，それほど重要であるとはいえない．

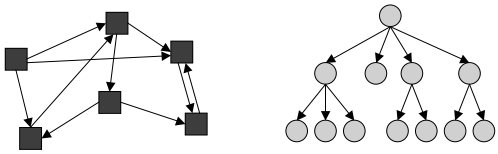

図 1　WWW(左) と Web サイト (右) のリンク構造

サイト内検索エンジンでは，具体的なコンテンツを持ち，容易に発見できるわけではないが，ある程度の数のページから参照されているようなページに大きな価値があると考えられる．

2.2 ハイパーリンク構造の相違

WWW のハイパーリンク構造は，Web サイトをひとつの節点とすると，規則性の少ない一般的な疎グラフである．このようなリンク構造に対して従来のスコアリング手法が有効であることは，WWW 検索エンジン Google の示すところである．

一方，Web サイト内のローカルなハイパーリンク構造は，トップページを根，具体的なコンテンツのあるページを葉とした木構造をベースとし，そこにいくつかのリンクを付加したグラフであると考えられる．

ここで，前節で議論した WWW 検索エンジンの検索要求は「根または根に近いページ」，サイト内検索エンジンの検索要求は「ある程度リンクが集中している葉または葉に近い内点」と，グラフ理論で定式化することができる．

このように，両者のハイパーリンク構造や検索要求の間には明白な違いがある．このことからも，PageRank などの WWW 検索エンジンで有効なスコアリング手法をそのままサイト内検索エンジンに適用するのは問題があることが推測できる．

2.3 従来の手法での問題点

Web サイトのハイパーリンク構造は木構造ベースとしており，サイト内の各ページは，ユーザがサイトを閲覧しやすいように単純な案内としての役割を持つトップページへのリンクや，親ページへのリンクを設定している場合が多い．

実際に，サイト内の全てのページがトップページや親ページへのリンクを持っていることも珍しくない．その結果，根に近づくほど多くのリンクを受けるような木構造となる．

このようなハイパーリンク構造に対して従来のスコアリングを行うと，トップページやその周辺のページに最も大きなスコアが割り当てられ，トップページから遠ざかるにつれてスコアが小さくなっていくことが予想される．

WWW 検索という視点で見れば，トップページやその周辺のページに高いスコアが割り当てられることは WWW 検索エンジンの精度向上に貢献しており，むしろ都合のよいことである．

しかし，サイト内検索という視点では，葉や葉に近い内点にある価値の高い情報を見逃している可能性があり，好ましい結果であるとは言い難い．

具体例として，Web サイトのリンク構造の概念図を図 2 に示す．各節点の値は，Web サイト内のローカルリンクからの被リンク数を表している．この値が直接 PageRank などの従来のスコアリング手法で割り当てられるスコアになるわけではないが，スコアを推測するもっとも単純な指標となる．

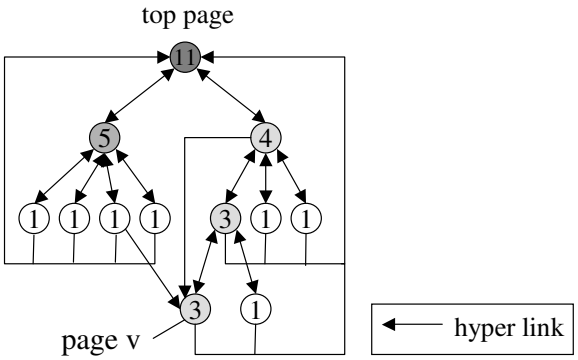

図 2　Web サイトのリンク構造の概念図

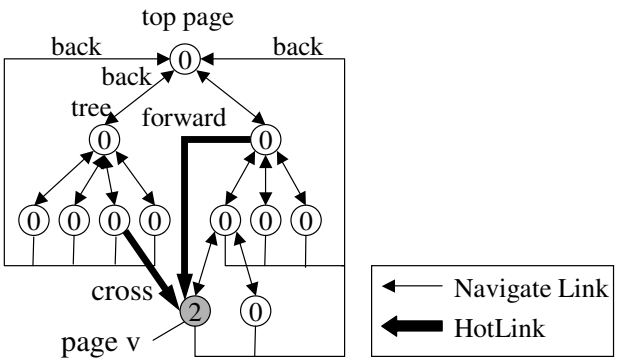

図 3　HotLink 法によるスコアリング

この例では，Web サイト内の全てのページがトップページと親ページへのリンクを持っている．これらのリンクによって，トップページに近づくほど多くのリンクを受けるようになっている．

その結果，リンク構造を見る限りこの Web サイトの中で価値の高そうなページ v が，トップページやその周辺のページに埋もれてしまっていることが分かる．

2.4　原因の考察

従来のリンク構造を利用したスコアリングでは，ハイパーリンクを推薦関係とみなして，張られているリンクによってそのページのスコアを決定する．

ここでページ作成者の視点に立って，あるページにリンクを張る行為の持つ意味を考えてみると，リンク先の情報を推薦または引用している場合と，ユーザがサイト内を閲覧しやすいように単純な案内の役割を持たせている場合があることが分かる．

トップページへのリンクは，ページ作成者がトップページにある情報を推薦または引用しているのではなく，ユーザの利便性を考慮していると考えるのが妥当である．

以上の考察により，WWW 検索エンジンでは有効な従来の手法がサイト内検索エンジンのスコアリング手法として有効に働かないのは，すべてのリンクを推薦関係としているところに原因があると考えられる．

よって，Web サイト内のすべてのリンクを推薦関係とみなすようなスコアリング手法は，サイト内検索エンジンにおいては好ましいとは言い難い．

3.　HotLink を用いたスコアリング

この問題を解決するために，あるページへリンクを張る行為の持つ意味に着目し，Web サイト内の全てのローカルリンクを単純な案内としての役割を持つ Navigate Link と，推薦または引用関係にある HotLink [1] の 2 種類に分類する．

HotLink を用いたスコアリングとは，Web サイト内の全てのローカルリンクを Navigate Link と HotLink に分類し，HotLink のみを用いてスコアリングを行う手法である．

ここでは，HotLink を用いたスコアリングとして HotLink 法を紹介し，HotLink 法を発展させた HL-PR 法を提案する．

3.1　リンクの分類方法

ここで，Web サイト内のすべてのリンクを Navigate Link と HotLink に分類する必要があるが，これを自動的に行う方法について考える．

Web サイトのハイパーリンク構造は，トップページを根とし，コンテンツのカテゴリごとに部分木を形成しているような木構造である．リンク構造からこの木構造を抽出することにより，Web サイト内のすべてのリンクは，その性質から，木を構成する tree edge，リンク先がリンク元の先祖である back edge，リンク先がリンク元の子孫である forward edge，それ以外の cross edge の 4 種類に分類することができる [3].

このうち，forward edge はユーザがすばやくその情報にアクセスできるようにリンクをたどる回数を減らす役割を持っており，リンク先の情報を推薦していると考えられる．

また，cross edge はある部分木から別の部分木へのリンク，すなわち，自分のページのカテゴリとは異なるカテゴリへのリンクなので，リンク先の情報を推薦または引用していると考えられる．

以上の理由から，Web サイト内のローカルリンクから木構造を抽出することによって決まる forward edge と cross edge を HotLink，tree edge と back edge を Navigate Link として定義する．

表 1　リンクの分類方法

リンクの種類	edge の種類	edge の分類方法
Navigate Link	tree edge	木を構成する edge
	back edge	リンク先がリンク元の祖先
HotLink	forward edge	リンク先がリンク元の子孫
	cross edge	otherwise

3.2　HotLink 法

HotLink 法 [4] は，HotLink の被リンク数をそのページのスコアとする，シンプルなスコアリング手法である．

先程の具体例（図 2）に HotLink 法を適用してスコアリングを行った結果を図 3 に示す．各節点の値は HotLink の被リンク数で，すなわち HotLink 法によるスコアである．

図 2 と比較すると従来のスコアリング手法との違いは明白で，従来の手法ではトップページやその周辺のページに埋もれてしまうようなページ v に，高いスコアを割り当てることに成功している．

この例は，説明のための実在しない Web サイトだが，次

は実在する Web サイトへの HotLink 法の適用例を図 4 に示す．対象としたのは，東北大学情報科学研究科外山研究室 (http://www.nue.riec.tohoku.ac.jp/) の Web サイトで，ページ総数は 43，リンク総数は 89 であった．

英語のトップページ A を根とする Shortest-Path Tree を tree edge と仮定してリンクを分類した．図においては，見易さのために back edge を除去してある．

各節点の値は HotLink の被リンク数，各節点に添えられている値は PageRank によるスコアを最大値が 100 となるように正規化した値である．また，特徴的なページにはアルファベットを付けて表 2 でページの簡単な説明を行っている．

PageRank によるスコアは，英語のトップページ A に最も高いスコアが割り当てられ，A の周辺のページに比較的高いスコアが割り当てられる傾向があることが分かる．

一方，HotLink 法によるスコアは，PageRank が中程度に高く，トップページから決して近くはないページの重要度を上げることがあることが分かる．

表 2　ページの簡単な説明

ページ	説明
A	英語のトップページ
B	日本語のトップページ
C	研究関連のリンク集
D	教授紹介の英語のトップページ
E	教授紹介の日本語のトップページ
F	助手紹介の英語のトップページ
G	助手紹介の日本語のトップページ

3.3　HL-PR 法

種々の Web サイトに対して PageRank によるスコアと HotLink 法によるスコアの比較実験を行ったところ，両手法におけるランキング上位のページが似た傾向を持つことがあった．これは，サイト内のほとんどすべてのページからリンクを受けているようなページ (トップページを除く) が存在するときに顕著であった．

PageRank が高いページは，WWW 検索の視点で見るとそのサイトを代表する興味深いページだが，サイト内検索の視点で見るとトップページから容易に発見できるようなページであることが多い．

そこで，HotLink と PageRank の値を最大値が等しくなるように正規化し，これらの値の差分をスコアとする手法，HL-PR 法を提案する．

差分をスコアとする直感的な理由は，PageRank を WWW 検索での重要度，HotLink をサイト内検索での重要度であるとして，PageRank が極端に高いページをカットすることにより，ランキングの改善を図ろうというものである．

種々の Web サイトで実験を行った結果，PageRank と HotLink はページ数が多くなると似たスコアの分布を示す傾向があることが分かった．よって，対数を取るなどの操作をすることなしに差分を取ることは，妥当な演算であると思われる．

3.4　木の抽出方法

Web サイトのリンク構造から木構造を抽出すれば HotLink が決定し，HotLink を用いたスコアリングを行うことができる．ここでは，Web サイトの木構造を適切に抽出する方法について議論する．

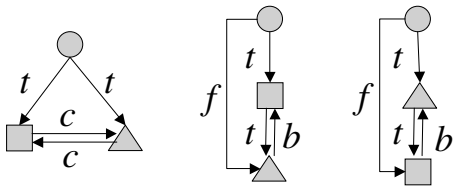

図 5　tree edge 決定の困難性

しかし，リンク構造のみから適切な木を選択するのは非常に難しい問題である．図 5 は Web サイトによく現れる部分構造である．図中の記号，t，b，f，c はそれぞれ tree edge，back edge，forward edge，cross edge を表している．

この部分構造において，tree edge の選び方は 3 通り考えられるが，このリンク構造を見る限りではどれが適切かを論じることはできない．

しかし，ここで Shortest-Path Tree を tree edge と仮定すると，比較的適切な木を選択できるのではないかと予想される．Shortest-Path Tree は幅優先探索によって得られる木で，他の候補の木に比べて幅が広く，高さが低い木になることから，Web サイトのリンク構造に近いと考えられるためである．

ただ問題となるのは，適切な木では forward edge となるリンクが，Shortest-Path Tree ではすべて tree edge として認識されてしまう点である．Web サイトの木構造を正確に抽出する，という観点ではあまり好ましい結果ではない．

しかしこの場合，適切な木では forward edge で指される node はかならず cross edge で指されることになる．forward edge がなくなることで正しい木構造からは崩れてしまうが，スコアリングのための前処理という観点で考えると，この性質は Shortest-Path Tree を仮定するとよいスコアリングを行える理由の 1 つになるのではないかと考えられる．

また，ディレクトリ情報を用いることによってより適切な木を抽出することができる可能性もあるが，本研究では Web サイトのローカルなリンク情報のみを用いてスコアリングを行うことを目標とする．

4.　実　　験

本研究では，種々の Web サイトに対して PageRank，HotLink 法，HL-PR 法によるスコアリングの比較を行い，提案手法である HL-PR 法の有効性について検証を試みている．

また，PageRank と HotLink 法によるスコアの分布を図示し，両手法におけるスコアの差分を取ることが妥当な演算であることを示している．

実験はすべて，CPU:PentiumIII 500[MHz]，Memory:128[MB]，OS:Windows2000 Professional という環境のもとで行った．GNU Wget を用いて Web ページの収集を行い，Perl を用いて収集した html ファイルに id を割り当て，PageRank 計算のための隣接行列や後述の LEDA 用グラフファイルを作成した．

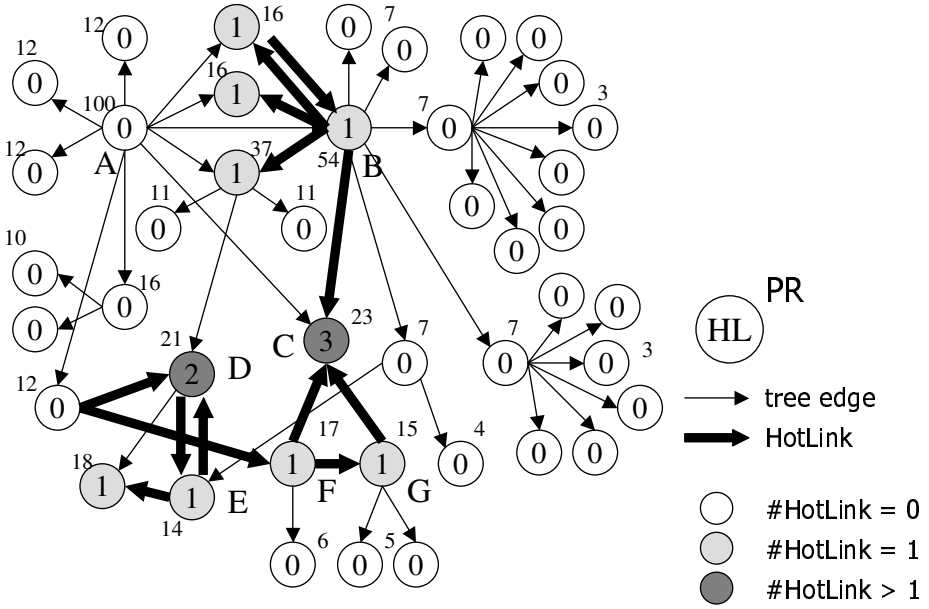

図 4　実在する Web サイトでの HotLink 法の適用例

また，GNU Octave を用いて隣接行列から PageRank を算出し，C++および C++のクラスライブラリである LEDA を用いてグラフに対する処理を行った．

以下，小規模 Web サイトとしてページ数が 100 程度の場合，大規模 Web サイトとしてページ数が 10000 程度の場合について実験結果を示す．

4.1　小規模 Web サイトでの実験結果

Windows.FAQ(http://winfaq.jp/) を対象として，PageRank，HotLink 法，HL-PR 法によるスコアリングの比較を行った．対象 Web サイトの総ページ数は 109，総リンク数は 1045 であった．ここでは，結果を見やすくするために，それぞれのスコアリングについて最大のスコアを 100 として正規化を行っている．

表 3　PageRank における上位のページ

HL	PR	HL-PR	URL(http://winfaq.jp/)
0	100	-100	index.html
0	29	-29	w2k/index.html
4	24	-20	whatsnew.html
6	24	-18	w98/index.html
13	24	-11	wme/index.html
4	23	-19	wxp/index.html
30	21	9	c/9xboot.html
38	19	19	c/network.html
63	19	44	w2k/boot.html
54	18	36	w2k/custom.html

PageRank によるスコアが上位のページを表 3 に，HotLink 法によるスコアが上位のページを表 4 に，HL-PR 法によるスコアが上位のページを表 5 に，スコアが下位のページを表 6 に示す．

PageRank におけるスコアが上位のページは，トップページやその周辺のページで占められる結果となった．これは，トップページへの膨大な back edge によるものであると考えられる．このようなページは，サイト内検索エンジンを利用するユーザにとってはほぼ既知であると思われる．

表 4　HotLink 法における上位のページ

HL	PR	HL-PR	URL(http://winfaq.jp/)
100	6	94	wme/network.html
100	8	92	wme/hints.html
96	5	91	sidenavi2.html
83	6	77	wme/custom.html
79	6	73	pinghowto.html
75	10	65	w2k/disk.html
75	2	73	remotedesktop.html
75	0	75	openwithnotepad.html
71	8	63	wme/pchealth.html
67	10	57	news.html

表 5　HL-PR 法によるスコアが上位のページ

HL	PR	HL-PR	URL(http://winfaq.jp/)
100	6	94	wme/network.html
100	8	92	wme/hints.html
96	5	91	sidenavi2.html
83	5	78	wme/custom.html
75	0	75	openwithnotepad.html
79	6	73	pinghowto.html
75	2	73	remotedesktop.html
75	10	65	w2k/disk.html
71	8	63	wme/pchealth.html
63	2	61	mp3.html

一方，HotLink 法，HL-PR 法におけるスコアが上位のページは，ほぼ同じ傾向となった．これらのページは具体的なコンテンツを持ち，ある程度リンクが集まっているページである．Web サイトの訪問者は，このようなページを検索するためにサイト内検索エンジンを利用するのではないかと予想される．

HL-PR 法によるスコアが最小となったのはトップページで，続いてトップページに近いページに低いスコアが割り当てられ

表 6　HL-PR 法によるスコアが下位のページ

HL	PR	HL-PR	URL(http://winfaq.jp/)
0	100	-100	index.html
0	29	-29	w2k/index.html
4	24	-20	whatsnew.html
4	23	-19	w98/index.html
4	23	-19	wxp/index.html
0	13	-13	wxp/network.html
0	13	-13	w2k/w2kfaq.html
0	13	-13	c/9xdisk.html
0	11	-11	w2k/ad.html
0	11	-11	customizetool.html

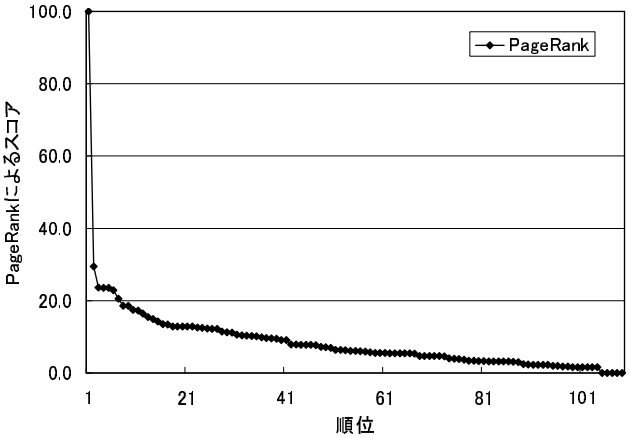

図 6　PageRank によるスコアの分布

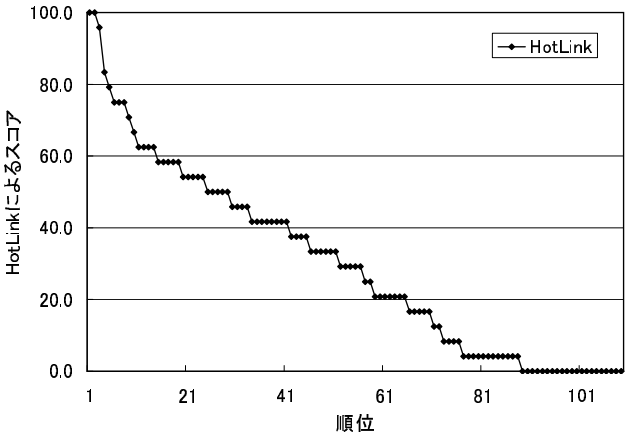

図 7　HotLink によるスコアの分布

る結果となった．これは，各ページからトップページへの膨大な back edge によってトップページの PageRank が極端に高くなり，トップページからリンクしているページがその影響を受けた結果，これらのページのスコアが低くなったと考えられる．

WWW という視点から見れば，この Web サイトで重要なページは PageRank における上位のページや，HL-PR 法における下位のページであるかもしれない．しかし，トップページまでたどり着いているユーザにとってこれらのページはほぼ既知であり，サイト内検索を利用して発見したいようなページではないと思われる．

また，PageRank と HotLink 法によるスコアの分布を図 6，7 に示す．この Web サイトにおいては，両手法におけるスコアの分布には明確な違いが現れた．しかし，大規模 Web サイトでの実験結果から，ページ数が増加するにつれて両者の分布が類似してくることが予想される．

4.2　大規模 Web サイトでの実験結果

@IT(http://www.atmarkit.co.jp/) の Web サイトを対象として，PageRank，HotLink 法，HL-PR 法によるスコアリングを行った．対象 Web サイトの総ページ数は 10240，総リンク数は 284168 であった．ここでも同様に，結果を見やすくするために，それぞれのスコアリングについて最大のスコアを 100 として正規化を行っている．

PageRank によるスコアが上位のページを表 7 に，HotLink 法によるスコアが上位のページを表 8 に，HL-PR 法によるスコアが上位のページを表 9 に，スコアが下位のページを表 10 に示す．

小規模サイトでの実験とは異なり，PageRank と HotLink 法によるスコアが上位のページが，同じ傾向となる結果が得られた．これは，対象 Web サイトのサーバ側ですべてのページの一部を集中管理しており，Web サイト内の大部分のページに同じリンクが設定されているためである．実際に確認を行った結果，これらのリンクは主要コンテンツのインデックスページや最新情報へのリンクであることが分かった．

しかし，ほとんどすべてのページからリンクされているとはいえ，例えばサイト管理者への問い合わせのページに最大のスコアが割り当てられてしまうのは，あまり好ましい結果とは言えない．

それに比べて HL-PR 法では，このページをカットすることに成功しており，ランキングの分布も全体的に改善されたような印象を受ける．

HL-PR 法の成果は，スコアが下位のページでも確認することができる．/aboutus 以下のページが多く見られるが，これらはサイト紹介やスタッフ紹介などのページである．

これらのページは，多くのページからリンクを受けてはいるものの，特定の情報を紹介するために外へリンクを張ることは少なく，これらのページ間では密なコミュニティが形成されている．その密なリンク構造によって PageRank が閉じこめられ，高い PageRank が割り当てられた結果，スコアが低くなったと考えられる．

また，PageRank と HotLink 法によるスコアの分布を図 8，9 に示す．両者の分布を比較して，類似した分布を持つと言ってよいだろう．HotLink 法が順位の後半で落ち込んでいるが，そのスコアの差は 0.1 未満であることから，誤差の範囲内である．

表 7　PageRank における上位のページ

HL	PR	HL-PR	URL(http://www.atmarkit.co.jp/)
0	100	-100	index.html
100	69	31	aboutus/contact_us/contact_us.html
90	44	46	applymember/club_index.html
93	38	55	aig/searchtop.html
89	28	61	scenter/learning/index.html
88	28	60	club/mail_news.html
88	26	58	scenter/job/index.html
15	25	-10	icd/index.html
86	23	63	ad/adindex/index/adindex.html
64	21	43	aboutus/sponsor/sponsor.html

表 8　HotLink 法における上位のページ

HL	PR	HL-PR	URL(http://www.atmarkit.co.jp/)
100	69	31	aboutus/countact_us/contact_us.html
93	38	55	aig/searchtop.html
90	44	46	applymember/club_index.html
89	28	61	scenter/liarning/index.html
89	26	63	scenter/job/index.html
88	28	60	club/mail_news.html
88	17	71	news/200212/20/oracle.html
88	19	69	news/200212/20/bea.html
88	19	69	news/200212/20/verisign.html
86	23	63	ad/adindex/index/adindex.html

表 9　HL-PR 法における上位のページ

HL	PR	HL-PR	URL(http://www.atmarkit.co.jp/)
88	17	71	news/200212/20/oracle.html
88	19	69	news/200212/20/bea.html
88	19	69	news/200212/20/versign.html
84	20	64	news/index.html
86	23	63	ad/adindex/index/adindex.html
88	26	62	scenter/job/index.html
89	28	61	scenter/learning/index.html
89	28	61	club/mail_news.html
78	20	58	news/keyword-news.html
64	7	56	fjava/devs/xpd01/xpd01.html

表 10　HL-PR 法における下位のページ

HL	PR	HL-PR	URL(http://www.atmarkit.co.jp/)
0	100	-100	index.html
1	17	-16	aboutus/staff/staff.html
1	17	-16	aboutus/press/press.html
1	16	-15	info/sitemap/sitemap.html
1	16	-15	aboutus/profile/prifile.html
3	16	-13	aboutus/termofuse/termofuse.html
3	16	-13	aboutus/index.html
1	14	-13	aboutus/p_policy/p_policy.html
1	14	-13	aboutus/copyright/copyright.html
1	14	-13	aboutus/b_policy/b_policy.html

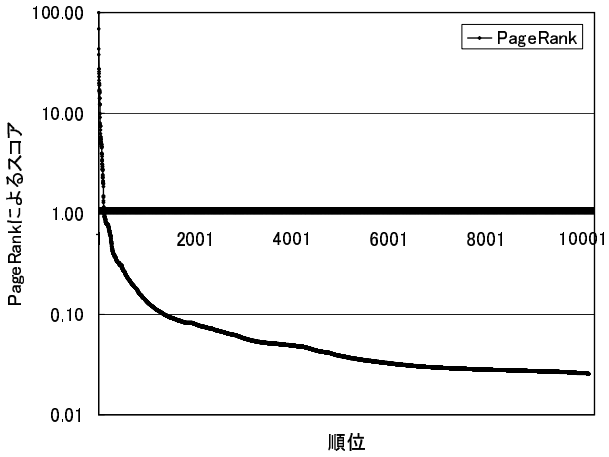

図 8　PageRank によるスコアの分布

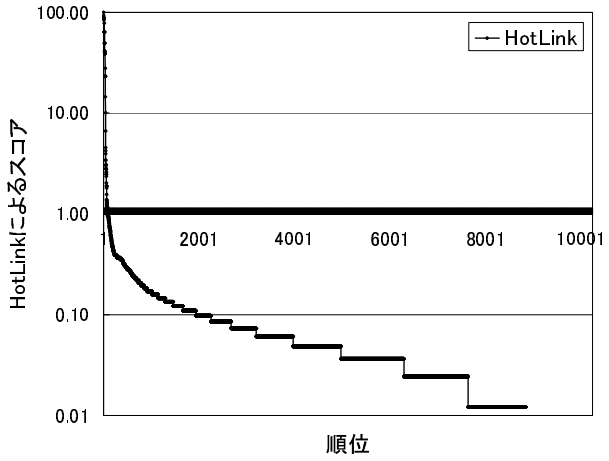

図 9　HotLink によるスコアの分布

5. 考　　察

本論文では，大規模 Web サイトにおいてほとんど全てのページからリンクされているようなページは，PageRank と HotLink 法の両手法で高いスコアが割り当てられてしまう，という問題点に着目し，両手法でのスコアの差分を取ることでランキングの改善を図っている．

また，ランキング改善のための別のアプローチは，Web サイトから木構造を正しく抽出することである．

Web サイトの規模が大きくなることで誤った木を抽出したときの誤差も大きくなり，膨大な back edge が HotLink として認識され，極端なスコアの偏りが生じてしまうことが考えられる．

また，そもそも Web サイトのローカルリンク構造は必ず単純な木構造になっているのか，という問題がある．

Web サイト内の大部分のページにまったく同じインデックスが設定されているのはよくあることだが，この場合，リンク構造としてはインデックスとしてリンクされているページでクリークが構成されることになる．また，Web サイト内の全てのページはクリークを構成するページへのリンクを持っている．

このようなリンク構造の概念図を図 10 に示す．節点の色は，濃いほど HotLink 法によるスコアが高くなることを示している．

クリーク中で探索を開始する節点は，クリークにぶら下がる

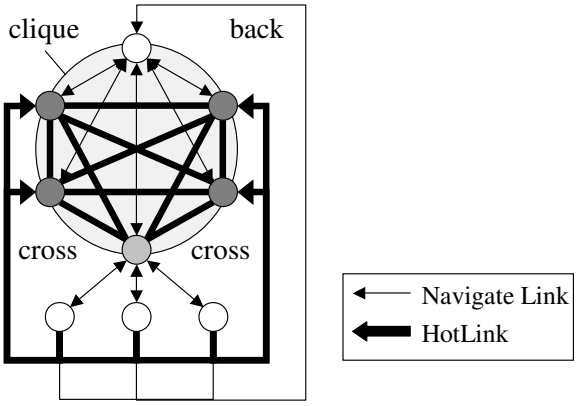

図 10　クリークのあるローカルリンク構造の概念図

ページ数が増えてもまったく HotLink が割り当てられないのに対して，その他の節点はクリークにぶら下がるページ数が増えるにつれて膨大な HotLink を受けることになり，極端なスコアの偏りが生じている．

これが，Web サイトの規模が大きくなると HotLink 法によるスコアが極端に高いページが現れる原因の一つではないかと考えられる．対策としては，クリークを一つの節点とみなしてリンクの分類を行うことが挙げられる．

6. ま と め

本論文では，サイト内検索エンジンのためのスコアリング手法として，Web サイトのリンク構造から木構造を抽出することによって決定する HotLink を Web ページのスコアリングに用いることを提案した．

また，具体的なスコアリング手法として HotLink 法を改良した HL-PR 法を提案し，その有効性について検証を試みた．

今後は，引き続き種々の Web サイトで実験を行い，提案手法の有効性を検証していこうと考えている．その他の課題としては，HotLink の重み付けやテキストマッチングとの連携，Web サイトのリンク構造から木を抽出するアルゴリズムの改良などが挙げられる．

文　献

[1] P. Bose, J. Czyzowicz, L. Gasieniec, E. Kranakis, D. Krizanc, A. Pelc, and M. Martin. Strategies for Hotlink Assignment. *Proceedings of ISAAC2000, Springer LNCS 1969*, pp.23-34, 2000.

[2] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. *Technical Report, Computer Science Department, Stanford University*, 1998.

[3] T. Cormen, C. Leiserson, R. Rivest and C. Stein. Elementary Graph Algorithms, Chapter 22 of *Intoroduction to Algorithms second edtion*(2001), 527-560.

[4] 伊川洋平，定兼邦彦．サイト内検索のためのスコアリング手法．FIT 情報技術レターズ，LD-2，2002.