DEWS2005 1C-o2

ホットリンク解析によるサイト内検索エンジンのランキング改善手法

大規模なWebサイトへの適用と考察

伊川　洋平†　　定兼　邦彦††

† 日本アイ・ビー・エム株式会社 東京基礎研究所
†† 九州大学 大学院システム情報科学研究院

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，ハイパーリンクを解析し，各 Web ページの重要度に応じてスコアを割り当てるスコアリング技術が重要視されている．WWW 検索エンジンのためのスコアリング手法は広く研究されているが，サイト内検索エンジンのためのスコアリング手法は未だ確立しておらず，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．この点に着目して提案された，ホットリンクを用いたスコアリング手法は，ユーザーが特定のコンテンツへ直接ジャンプできるように設定するホットリンクを分析することで，重要なコンテンツページに高スコアを割当て，サイト内検索エンジンのランキングを改善する手法である．本手法は，これまで小規模な Web サイトにおいて有効性が検証されているが，大規模な Web サイトについては充分検証が行われていなかった．本論文では，本手法が大規模な Web サイト対しても，重要なコンテンツページに高スコアを割当てることを実験により確認し，サイト内検索エンジンにおいて有用なスコアリング手法であることを検証する．

キーワード　Web 構造解析，Web マイニング，情報検索

A Method for Improving the Ranking Results of Local-Web Search-Engines using HotLink Analysis

Application to Large-Scale Websites and Disscussions

Yohei IKAWA† and Kunihiko SADAKANE††

† IBM Research, Tokyo Research Laboratory
†† Department of Computer Science and Communication Engineering, Kyushu University

Abstract　Web-page scoring is a method to evaluate output from search engines by assigning a score to each page according to its importance. Google's PageRank algorithm is an efficient scoring method for WWW search-engines. However it is not suitable for searching local Webs because valuable information in the hyperlink structure of Web-pages is not utilized. Ikawa and Sadakane proposed scoring methods using HotLink with this point of view, and examined that our new scoring method captures important pages in a small-scale local Web. In this paper, we examine its practical effiectiveness in a large-scale local Web.

Key words　Link Analysis, Web Mining, Information Retrieval

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，多くのユーザーは Google (注1) (注2) のような Web 検索エンジンを日常的に利用している．

Web 検索エンジンの利便性を向上させる手段として，各 Web ページの重要度に応じてスコアを割り当てる技術，すなわち，Web ページのスコアリングが重要視されている．Web ページのスコアリングを用いて，ユーザー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手法は大別すると，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 [8], [9] に分類できる．本研究で扱うのは，後者のハイパーリンク構造を利用したスコアリングである．

(注1)：http://www.google.com/
(注2)：$Google^{TM}$ は Google Inc. の登録商標である

1.1 ハイパーリンク構造を利用したスコアリング

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されるため，不正にスコアを上げることが難しく，さらに，テキスト解析によるスコアリングと組み合わせることによって，信頼性の高い検索エンジンを構築することができる．

ハイパー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は PageRank (注3) [9] である．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 グラフ上の参照関係をランダムウォークとして定式化し，単純マルコフ過程を用いて各ページの滞留確率を計算し，スコアとして利用する手法である．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．

1.2 サイト内検索エンジンのためのスコアリング

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，ユーザー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．また，イントラネットの発展により，既存の WWW 検索エンジンが利用できない，社内 Web のようなインターナルな Web サイトに対する検索の重要性が高まっている [3].

サイト内での Web ページのリンク関係は，グラフ理論的に WWW とは異なり，有向木を根幹に持つ特殊なグラフとしてモデル化される．このモデルにおいては，WWW 検索エンジンにおいて有効な手法として知られる PageRank アルゴリズムは有効に働かず，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，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として，ホットリンクと呼ばれる特殊なリンクに注目する．Web におけるホットリンクとは，ユーザーが特定のコンテンツへ直接ジャンプできるように設定するリンクのことである．通常は，異なる Web サイトへ張られることが多く，リンク集として日常的によく見かけるものである．

本研究では，実際の Web サイトにおいても，ベースとなる構造に添加された，ユーザーの利便性のために張られたリンクが存在するということに着目する．例えば，ニュースサイトでは，トップページにおける最新ニュースへのリンク，あるニュースに関連の深い過去のニュースへのリンクがそれにあたる．このホットリンクを分析することで，重要なコンテンツページに高スコアを割当て，ランキングを改善するのがホットリンクを用いたスコアリング手法である．

(注3)：$PageRank^{TM}$ は Google Inc. の登録商標である

1.3 本論文の目的

ホットリンクを用いたスコアリング手法には，サイト内のローカルリンクからホットリンクを抽出するために，ランダムに固定した木を用いる HotLink 法 [4]，全ての最短路木を考えたときの平均の HotLink スコアを用いる HLAve 法 [6]，HotLink スコアと PageRank スコアの差分を取る HL-PR 法 [5] がある．

これらの手法は，これまで数百ページ程度の小規模な Web サイトにおいて実験が行われ，サイト内検索エンジンで有効なスコアリング手法であることが検証されている [6]．ところが，数万ページのオーダーの大規模な Web サイトに対する有効性は，充分検証が行われていなかった．本論文では，これらの手法が大規模な Web サイト対しても，重要なコンテンツページに高スコアを割当てることを実験により確認し，ホットリンクを用いたスコアリングが，サイト内検索エンジンにおいて有用な手法であることを検証する．

2. サイト内検索における PageRank の問題点

PageRank アルゴリズムは，WWW 検索エンジンでは有効なスコアリング手法だが，そのスコアリング結果は，サイト内検索エンジンのユーザーにとって好ましくないものである．ここでは，サイト内検索エンジンに PageRank スコアを用いる問題点を指摘し，その原因について考察する．

2.1 Web サイトのリンク構造の特徴

Web サイト内のページは，階層構造をもつディレクトリによって管理されていることから，トップページを根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．また，サイト内の各ページは，ユーザーがサイトを閲覧しやすいように，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．その結果，Web サイトのリンク構造は，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．

［観察 1］ 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た木構造をベースとしている．各ページは祖先のページへのリンクを持っていることが多く，根に近づくほど多くのリンクを受けるような構造となる．

2.2 PageRank の問題点

観察 1 に基づいた Web サイトのリンク構造に対して PageRank で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特に高いスコアが割当てられ，トップページから遠ざかるにつれてスコアが低くなっていく．

WWW 検索の視点では，Web 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．実際，トップページを見出すことにより，そこから求めたい情報に行き着けることは多いため，PageRank のランキングは WWW 検索エンジンのユーザーにとって好ましい．

一方，サイト内検索エンジンのユーザーは，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．従って，サイト内検索エンジンのユーザーにとっては，トップページやトップペー

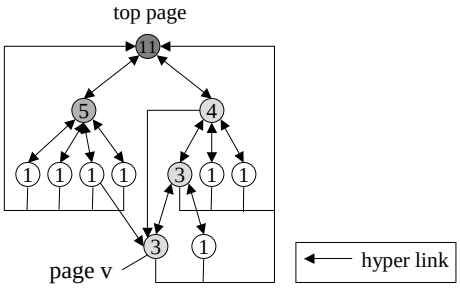

図 1　Web サイトのリンク構造の具体例

ジから容易に発見できるページは重要ではなく，ある程度の数のページから参照されているようなコンテンツページが重要である．

Web サイトのリンク構造の具体例を図 1 に示す．この Web サイトの各ページは，全ての祖先のページへのリンクを持っており，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な構造になっている．

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．PageRank でスコアリングを行うと，この値が大きいページに高いスコアが割り当てられることになる．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．

2.3 考　　察

PageRank アルゴリズム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，Web サイト内の全てのリンクを用いてスコアリングを行っているところに原因があると考えられる．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すことによってページの質を決定するが，Web サイト内の全てのリンクがリンク先のページを推薦しているわけではない．

実際に，ページ作成者があるページへリンクを張るとき，リンク先のページを推薦する場合だけでなく，ユーザー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．例えば，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を推薦しているのではなく，Web サイト閲覧の利便性を考慮してリンクを設定したと考えるのが妥当である．

従って，Web サイト内の全てのリンクを用いるのではなく，推薦関係にあるリンクを推定し，これらのリンクのみを用いてスコアリングを行うことを考える．ここで，推薦リンクを推定するために，Web におけるホットリンクに着目する．

2.4 Web におけるホットリンク

Web におけるホットリンクとは，ユーザーが特定のコンテンツへ直接ジャンプできるように設定するリンクのことである．通常は，異なる Web サイトへ張られることが多く，リンク集として日常的によく見かけるものである．

ホットリンクを Web サイト内のポピュラーなページへ添加することによって，トップページから目的のページにたどり着くまでの検索ステップ数を効率化することができる．計算理論分野では，どのようにホットリンクを張ると最も効率を向上できるかを考察した理論として，ホットリンク添加問題 [1] が研究されている．ホットリンク添加問題を考察すると，ホットリンクとページの重要性には大きな相関があることが分かる．従って，ホットリンクを用いてページの重要性を測ることができる．

本論文で紹介する HotLink 法 [4] は，与えられた Web サイトのリンク構造からホットリンクの集合を推定し，これを推薦関係にあるリンクとして，スコアリングに利用することを目標とした手法である．

3. HotLink を用いたスコアリング

HotLink を用いたスコアリングは，Web サイト内の全てのローカルリンクを，推薦関係にあるリンクを HotLink，単純な案内としての役割を持つリンクを Navigate Link と定義して，HotLink のみをスコアリングに利用する手法である．しかし，この定義はあいまいであり，また，Web グラフの構造のみからは判定できない．従って，より数学的な新しい定義を与え，上記のものを代替する．

また，[7] のように，リンクの種類を分類するために，テキストやタグなどのコンテンツ情報やレイアウト情報を利用することも考えられるが，本研究では，これらの情報を用いず，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．

3.1 リンクの分類方法

Web サイトの Web グラフ $G = (V, E)$ は，トップページに対応する節点 r が固定されており，r と各ページの間には必ず有向パスが存在している．また，トップページから各ページへのパスを併合すると全域木となる．

Web サイトのリンク構造は，観察 1 で述べたように，トップページを根とした木構造をベースとし，各ページが祖先のページへのリンクを持っていることが多い構造である．Web サイトのリンク構造から，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，次の 4 種類に分類することができる [2].

- 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

tree-edge は，Web サイトのベースとなる木構造を規定するものであるから，リンクに推薦の意味はなく，Navigate Link とするのが妥当である．

back-edge は，トップページや親ページへのリンクがこれにあたるが，2.3 節で述べたように，これらのリンクは Web サイト閲覧の利便性を考慮して設定されたと考え，Navigate Link とするのが妥当である．

forward-edge は，トップページから目的のページへたどり着くためのステップ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ため，HotLink とするのが妥当である．ニュースサイトの例では，トップページにおける最新ニュースへのリンクが forward-edge にあたる．

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

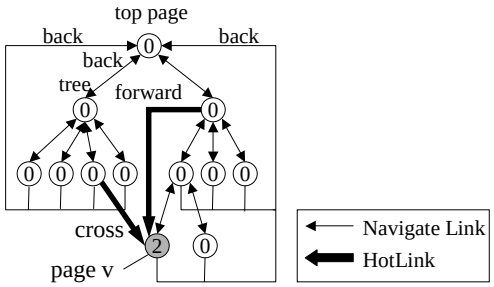

図 2　HotLink 法によるスコアリング

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ため，HotLink とするのが妥当である．ニュースサイトの例では，あるニュースに関連の深い過去のニュースへのリンクが cross-edge にあたる．

従って，HotLink と Navigate Link の分類方法を次のように定義する．

［定義 1］ Web サイトの Web グラフ G において，サイトのベース構造である，トップページ r を根とする G の全域木の辺を tree-edge とする．このとき，forward-edge と cross-edge を HotLink，tree-edge と back-edge を Navigate Link とする．

この分類により，2.2 節において，PageRank の問題点として指摘されていたトップページや祖先のページへのリンクは Navigate Link となり，スコアリングの対象外のリンクとして刈り込むことができる．

表 1　リンクの分類方法

リンクの種類	edge の種類	edge の定義
Navigate Link	tree-edge	木を構成する edge
	back-edge	リンク先はリンク元の祖先
HotLink	forward-edge	リンク先はリンク元の子孫
	cross-edge	otherwise

3.2　HotLink 法

HotLink 法 [4] は，HotLink の被リンク数を各ページのスコアとするスコアリング手法である．

［定義 2］ 節点 v に対して，$N^+(v) = \{e : v$ を終点に持つ有向辺 $\}$ とする．

［定義 3］ Web グラフ $G = (V, E)$ において，ページ $v \in V$ の HotLink スコア $HL(v)$ を次のように定義する．

$$HL(v) = \sum_{e \in N^+(v)} \delta(e)$$

ここで，$\delta(e)$ は辺 e が HotLink ならば 1，それ以外は 0 である．

先程のリンク構造の具体例（図 1）に HotLink 法を適用してスコアリングを行った結果を図 2 に示す．各節点の値は HotLink の被リンク数，すなわち HotLink スコアである．図 1 と比較すると PageRank スコアとの違いは明白で，PageRank 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．

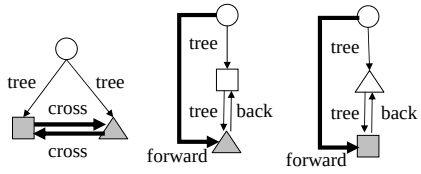

図 3　抽出する木構造のの候補

3.3　適切な全域木の抽出方法

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．しかしながら，木構造はあらかじめ与えられてはいない．従って，Web サイトの木構造を適切に抽出する方法の開発が必要である．

リンク構造のみから Web サイトの木構造を一意に決定することは，一般には不可能な問題である．図 3 は Web サイトによく現れる部分構造である．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．

従って，何らかの基準によって木構造の推定を行う必要がある．本論文では，高さが低い木が適切な木であると仮定し，最短路木を用いることを考える．最短路木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．

ここで注意すべきことは，適切な木では forward-edge となるリンクが，最短路木ではすべて tree-edge として認識されてしまう点である．従って，正確な木構造の抽出法とはなり得ない．しかしこの場合，適切な木では forward-edge で指される節点は必ず cross-edge で指されることになる（補題 3.1）．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 HotLink の消滅は生じない．実際，HotLink スコアの和を最大にする木として，必ず forward-edge が存在しない木を選べることが証明される（証明は省略）．代表的な forward-edge の存在しない木は最短路木であるので，この性質は最短路木を仮定するとよいスコアリングを行えるという一つの裏付けとなる．

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．

［補題 3.1］ 節点 u から v を指す forward-edge h を tree-edge に変更すると，節点 v を指していた tree-edge e は cross-edge になる．

［証明］ forward-edge の定義より，v は u の子孫であるから，u から tree-edge のみをたどって v へ到達する経路が存在し，その距離は 2 以上である．その経路上の v の直前の節点を w とする．ここで h を forward-edge から tree-edge に変更すると，e は tree-edge ではなくなり，w から見た v は先祖でも子孫でもない．よって定義により，e は cross-edge となる．

4. HotLink 法の改良

4.1 HLAve 法

HotLink 法は，ランダムに選んだ最短路木を一つ固定するため，固定する木が変化するとスコアリングの結果も変化してしまうという問題があった．HLAve 法は，全ての最短路木を考えたときの平均の HotLink スコアを用いることによって，この問題を解決する．

4.1.1 定 式 化

各辺に対して重み $w(e)$ を定め，HotLink の平均スコア $HLAve(v)$ を次のように定義する．

［定義 4］ 与えられた Web グラフ $G=(V,E)$ から最短路木を一様にランダムに 1 つ固定したとき，$e \in E$ が HotLink になる確率を $w(e)$ とし，ページ $v \in V$ のスコア $HLAve(v)$ を

$$HLAve(v) = \sum_{e \in N^+(v)} W(e) \cdot w(e)$$

とする．ここで，$W(e)$ は木の選び方に無関係な辺 e の重みを表す．

ここで $W(e)=1$ ならば，$HLAve(v)$ は全ての最短路木に対する $HL(v)$ の平均となる．次に，全ての $e \in E$ に対して $w(e)$ を求めるために $\tilde{G}$ を定義する．

［定義 5］ $G=(V,E)$ から任意の最短路木を構成したとき，tree-edge となりうる edge の集合を $\tilde{E}$ とし，$\tilde{G}=(V,\tilde{E})$ とする．

単純に $HLAve(v)$ を求めるならば，宇野 [10] による列挙アルゴリズムを用いて全ての最短路木の列挙を行い，それぞれの木に対して $HL(v)$ を求めて平均を計算すればよい．しかし，一般的に最短路木の数は膨大であるため，木の列挙は困難である．

［補題 4.1］ $\tilde{G}$ における最短路木の数を N，$\tilde{G}$ の根にあたる節点を r，v における $e \in \tilde{E}$ の入次数を $indeg_{\tilde{G}}(v)$ とすると，

$$N = \prod_{v \in V \setminus \{r\}} indeg_{\tilde{G}}(v)$$

が成立する．

［証明］ [6]

補題 4.1 より，最短路木の数 N は $|V|^{\Omega(|V|)}$ であり，最短路木を列挙して $w(e)$ を求めるのは難しいことが分かる．よって以下では，木の列挙を行わずに $w(e)$ を求める手法について検討する．

4.1.2 アルゴリズム

始めに，入力された Web グラフ G から $\tilde{G}=(V,\tilde{E})$ を構成する．具体的には，G において根 r を始点とした幅優先探索を行いながら，tree-edge となり得る edge を $\tilde{E}$ の要素に加えていくことで容易に $\tilde{G}$ を構成することができる．なお，幅優先探索を行う際には，節点 $v \in V$ に対して，r からの距離 $dist(v)$ と，$e \in \tilde{E}$ の入次数 $indeg_{\tilde{G}}(v)$ を求めておく．

$\tilde{G}$ と各節点に対する $dist(v)$，$indeg_{\tilde{G}}(v)$ を求めるアルゴリズム $Create\_\tilde{G}(G,r)$ を以下に示す．

［アルゴリズム 4.2］ $Create\_\tilde{G}(G,r)$

```
1  V ← {r}, i ← 0, Ẽ ← ∅
2  while V ≠ ∅
3      V_c ← ∅
4      for each v ∈ V
5          Mark v as visited
6          dist(v) ← i
7          for each v_c ∈ v に隣接する未訪問の節点集合
8              indeg_G̃(v_c) ← indeg_G̃(v_c) + 1
9              V_c ← V_c ∪ {v_c}, Ẽ ← Ẽ ∪ {(v, v_c)}
10     V ← V_c, i ← i + 1
```

$\tilde{G}$ を構成したら，以下のような場合分けによって，全ての $e=(s,t) \in E$ に対して，木の列挙を行わずに $w(e)$ を求めることができる．

［補題 4.3］ $e=(s,t)$ に対する重み $w(e)$ は次のように計算することができる．

$$w(e) = \begin{cases} 1 - \frac{1}{indeg_{\tilde{G}}(t)} & \text{if } e \in \tilde{E} \\ 1 & \text{if } e \in E - \tilde{E} \text{ and } dist(s) = dist(t) \\ 1 - p(t,s) & \text{if } e \in E - \tilde{E} \text{ and } dist(s) > dist(t) \\ 0 & \text{if } e \in E - \tilde{E} \text{ and } dist(s) < dist(t) \end{cases}$$

ここで，$p(v,w)$ は節点 v と w の間に有向パスが存在する確率を表す．

［証明］ [6]

4.1.3 $p(v,w)$ の計算

すべての $v,w \in V$ の組合せに対する $p(v,w)$ は，動的計画法を用いて $O(k|\tilde{E}|)$ で求めることができる．ここで，k は根 r からの距離 $dist(v)$ の最大値をあらわしている．

［補題 4.4］ w を固定し，v を w から先祖の方向へ移動させる，逆幅優先探索を考える．このとき，$V_c(v)$ を節点 v の子節点の集合であるとすると，次の関係が成立する．

$$p(v,w) = \begin{cases} 1 & \text{if} \quad v = w \\ \sum_{u \in V_c(v)} \frac{p(u,w)}{indeg_{\tilde{G}}(u)} & otherwise \end{cases}$$

［証明］ [6]

この関係を $v=r$ となるまで再帰的に適用することにより，固定された w に対する $p(v,w)$ を求めることができる．よって，w を動かすことにより，すべての $v,w \in V$ の組合せに対して $p(v,w)$ を求めることができる．

［補題 4.5］ 全ての $v,w \in V$ の組合せに対する $p(v,w)$ は $O(|V||\tilde{E}|)$ で計算できる．

［証明］ [6]

［定理 4.6］ 全ての節点 v に対する $HLAve(v)$ を求めるための計算量は $O(|V||\tilde{E}|)$ である．

［証明］ [6]

4.2 HL-PR 法

Web サイトによっては，サイト内のほぼ全てのページに同じリンクが設定されていることがある．その結果，トップページでないにもかかわらず，サイト内のほぼ全てのページからリンクされるページが存在することがある．このようなページを，本論文では intensely-linked page と呼ぶことにする．

［定義 6］ Web サイト内のほぼ全てのページからリンクされているページを intensely-linked page と呼ぶ．

複数の intensely-linked page がある Web サイトに対して，HotLink 法によるスコアリングを行うと，トップページ以外の intensely-linked page へのリンクの刈り込みは有効に行われず，極端に高いスコアが割当てられてしまう．

そこで，HotLink スコアと PageRank スコアの差分を取り，PageRank が極端に高いページをカットするのが HL-PR 法である [5]．intensely-linked page は PageRank スコアも極端に高いため，HL-PR 法によってランキングを改善することができる．

HL-PR スコアは，次のように定義される．

［定義 7］ HotLink スコアと PageRank スコアを最大値が等しくなるように正規化したとき，これらの値の差分を HL-PR スコアとする．

HLAve 法に対しても同様に，HLAve-PR スコアを定義する．

［定義 8］ HLAve スコアと PageRank スコアを最大値が等しくなるように正規化したとき，これらの値の差分を HLAve-PR スコアとする．

表 2　NHK の Web サイトで PageRank スコアが上位のページ

HL	HA	PR	HA-PR	URL
0	0	100	-100	index.html
7	8	16	-8	tv50/index.html
2	2	16	-14	hensei/index.html
10	11	16	-5	tv50menu/menu.html
5	5	13	-8	englishtop/index.html
4	4	13	-9	tvnavi/pal1/startchanet.html
2	2	13	-11	omoban/index.html
73	73	11	62	toppage/privacy.html
0	0	11	-11	strl/index.html
51	51	10	41	toppage/nhk_info/copyright.html

表 3　NHK の Web サイトで HLAve スコアが上位のページ

HL	HA	PR	HA-PR	URL
100	100	6	94	sapporo/pg_guide.html
97	97	4	93	sapporo/hokuhoku/index.html
73	73	11	62	toppage/privacy.html
51	51	10	41	toppage/nhk_info/copyright.html
47	47	3	44	sch/help/index.html
46	46	2	44	partner/faq/index.html
46	46	2	44	partner/boshuu/index.html
46	46	2	44	partner/yotei/yotei_1.html
46	46	2	44	partner/kurashi/index.html
42	42	2	40	bunken/book-jp/b4-j.html

5. 実　　験

本研究では，ページ数が数万のオーダーの大規模な Web サイトに対して，PageRank，HotLink 法，HLAve 法，HL-PR 法によるスコアリングの比較実験を行い，大規模な Web サイトに対する HotLink を用いたスコアリング手法の有効性を検証した．

結果の表において，PR，HL，HA，HA-PR はそれぞれ PageRank スコア，HotLink スコア，HLAve スコア，HLAve-PR スコアを表しており，各スコアは，それぞれの手法における最大のスコアを 100 として正規化を行ったものである．また，Indeg は各ページにおけるハイパーリンクの入次数を，HLAve は HLAve 法により求めた HotLink の平均入次数，すなわち，正規化を行っていない HLAve スコアを表している．

5.1 NHK の Web サイトでの実験結果

NHK の Web サイト(注4)での実験結果を示す．このサイトの総ページ数は 99,743，総リンク数は 466,983 であった．ページあたりの，同じサイト内のページへの平均リンク数は 4.7 で，比較的疎な構造の Web サイトといえる．実験結果を表 2 から表 5 に示す．

5.1.1 PageRank スコアの問題点

表 2 より，PageRank は，トップページ付近の index ページに高いスコアを割り当てていることが分かる．これらのページは，WWW 検索においてはこの Web サイトを代表する重要なページだが，この Web サイトの訪問者にとっては，容易に発見できるあまり重要ではないページである．

表 4　NHK の Web サイトで Indeg-HLAve の値が大きいページ

Indeg	HLAve	IN-HA	URL
10988	0	10988	index.html
1928	4	1924	sapporo/index.html
1549	8	1541	strl/index.html
1494	1	1493	sapporo/hokuhoku/h_bussan.html
853	106	747	bunken/index.html
785	5	780	school/index.html
736	102	634	midnight/stock/index.html
578	4	574	ganko/index.html
699	217	483	nagano/index.html
461	11	450	strl/publica/dayori-new/index-e.html

表 5　NHK の Web サイトで HotLink スコアと HLAve スコアの差分が大きいページ

HL	HA	PR	|HL-HA|	URL
2	30	2	28	partner/index.html
0	20	1	20	sports/baseball/index.html
40	20	1	20	sports/mlb/index.html
46	27	2	19	partner/sitemap/index.html
7	27	2	20	partner/p_guest/index.html
34	17	5	17	miyazaki/jouhou/index.html
46	31	2	15	partner/cooking/index.html
24	9	1	15	bunken/index-e.html

（注4）：http://www.nhk.or.jp/

また，ランキングの上位に各ディレクトリの *index.html* が多いことから，PageRank は，各トピックを構成する部分木の根にあたる index ページに，より高いスコアを割り当てる傾向があることが分かる．その理由は，同じトピックのページ群の中では index ページが最大スコアとなるトップページに最も近いということと，子孫のページからの Navigate Link によるものであり，そのページを推薦する意図のリンクが集中した結果ではないことが多い．その結果，index ページの下にある重要なコンテンツページが埋もれてしまっている可能性がある．

5.1.2 HotLink を用いたスコアの有効性

一方，HotLink 法や HLAve 法 (表 3) では，トップページやトップページに近い index ページへの Navigate Link を刈り込み，これらのページのスコアを抑えることによって，相対的にコンテンツページの重要度を強調することに成功している．ここで上位に現れているページは，トップページからは直接リンクされていない各番組のページや，サイト内の各ページから高頻度で引用されている，プライバシーポリシーや著作権について述べられているページである．

また，HLAve 法によって刈り込まれたリンクの数，すなわち，Indeg-HLAve の値が大きい順にソートした結果を，表 4 に示す．この表に現れているページのほとんどが index ページであることから，実際に index ページが多くの Navigate Link からリンクされており，HLAve 法を適用することによって，これらの Navigate Link をうまく刈り込めていることが確認できる．

5.1.3 HL スコアと HLAve スコアの比較

表 5 は，HL スコアと HLAve スコアの差分が大きいページを示している．これより，ランダムに選んだ最短路木を用いる HL スコアは，効率よく Navigate Link を刈り込むように木を選択することもあるが，同様に，適切でない木を選択することによって，HL スコアが増大してしまうこともある．一方，HLAve スコアは入力グラフに対して一通りに決まる値であるため，HLAve スコアの方が頑健な値であることが分かる．

5.2 Web サイト@IT での実験結果

Web サイト@IT(注5)での実験結果を示す．このサイトの総ページ数は 17,018，総リンク数は 558,298 であった．ページあたりの，同じサイト内のページへの平均リンク数は 32.8 で，NHK の Web サイトとは対照的に，非常に密な構造の Web サイトといえる．これは，サイト内のほぼ全てのページに設定されている，トップページや親ページ，メイントピックの index ページへのリンクによるものである．なお，結果の表において，URL の先頭に * がついているページは，Web サイト内の 90% 以上のページからリンクされている intensely-linked page である．実験結果を表 6 から表 10 に示す．

5.2.1 HotLink を用いたスコアリング手法の問題点

表 6 より，PageRank は，トップページ付近の index ページや，intensely-linked page に極端に高いスコアを割り当てており，サイト内検索エンジンにとって好ましい結果ではないことが分かる．

また，表 7 より，HotLink 法や HLAve 法は，一部の PageRank スコアの高いページを刈り込むことができているように見えるが，実際にはこれらのページへの Navigate Link はほとんど刈り込まれていない．よって，このランキングの変化は，Navigate Link を刈り込んだ結果ではなく，純粋にハイパーリンクの入次

表 6 Web サイト@IT で PageRank スコアが上位のページ

HL	HA	PR	HA-PR	URL
0	0	100	-100	*index.html
98	99	98	1	*fbiz/index.html
100	100	90	10	*aboutus/contact_us/contact_us.html
71	72	73	-1	job/index.html
50	50	61	-11	aboutus/termofuse/termofuse.html
80	80	45	-35	job/jc/index.html
1	1	40	-39	aboutus/index.html
1	1	38	-37	aboutus/p_policy/p_policy.html
80	80	36	44	ad/adindex/index/adindex.html
57	57	28	29	club/mail_news.html

表 7 Web サイト@IT で HotLink スコア，HLAve スコアが上位のページ

HL	HA	PR	HA-PR	URL
100	100	90	10	*aboutus/contact_us/contact_us.html
98	99	98	1	*fbiz/index.html
80	80	45	35	job/jc/index.html
80	80	36	44	ad/adindex/index/adindex.html
76	76	25	51	news/200412/17/hp.html
76	76	25	51	news/200412/18/symantec.html
76	76	25	51	news/200412/17/miracle.html
76	76	25	51	news/200412/17/oracle.html
76	76	25	51	news/200412/18/cisco.html
76	76	26	50	news/200412/18/sun.html

表 8 Web サイト@IT で Indeg-HLAve の値が大きいページ

Indeg	HLAve	IN-HA	URL
16714	0	16714	*index.html
12612	9495	3117	news/keyword-news.html
7321	5871	1450	info/sitemap/sitemap.html
2278	1285	993	fwin2k/index.html
1953	1094	859	icd/index.html
1848	1163	685	flinux/index.html
2362	1698	664	fdotnet/index.html
1626	1167	459	fsys/index.html
1321	980	341	fjava/index.html
13608	13293	314	ad/adindex/index/adindex.html

表 9 @IT の Web サイトで HotLink スコアと HLAve スコアの差分が大きいページ

HL	HA	PR	|HL-HA|	URL
31	35	27	4	info/sitemap/sitemap.html
5	7	6	2	icd/index.html
98	99	98	1	*fbiz/index.html
1	2	1	1	fwin2k/win2ktipsindex.html
8	7	4	1	fsys/index.html
11	10	5	1	fdotnet/index.html

(注5)：http://www.atmarkit.co.jp/

数に依存した結果となっている．

ここで，この Web サイトに対する HLAve 法の効果を確認するために，HLAve 法によって刈り込まれたリンクの数が大きい順にソートした結果を表 8 に示す．これより，HLAve 法を適用することによって，index ページへの Navigate Link をある程度刈り込むことに成功してはいるものの，疎な構造の Web サイトと比べて（表 4），あまり効果的にリンクを刈り込めていないことが分かる．

疎な構造の Web サイトでの結果と異なるもうひとつ特徴としては，HL スコアと HLAve スコアの間にあまり差が無いことである（表 9）．これは，どの木を選んでも効果的に Navigate Link を刈り込めないことを示しており，この Web サイトの構造が観察 1 で述べられているように，先祖のページのみに Navigate Link が張られるのではなく，先祖以外の特定のページにも多くの Navigate Link が張られていることを示している．

以上の結果から，Navigate Link を効果的に刈り込めない原因は，サイト内に Navigate Link が密に張り巡らされており，Web サイトが観察 1 で述べられているような単純な構造ではないためであると考えられる．

5.2.2 HL-PR 法の効果

このような，密な構造を持つ Web サイトに対しては，HL-PR 法が有効である．表 10 に，HLAve-PR スコアでソートした結果を示す．HLAve-PR スコアは，トップページ付近の intensely-linked page を完全に刈り込み，$news/index.html$ のスコアを抑え，$news/$ 以下のコンテンツページの重要度を強調することに成功している．また，この表に現れていない部分でも，index ページのスコアが抑えられ，相対的にコンテンツページのスコアが高くなり，全体的にランキングが改善されている．

このような結果が得られる原因としては，性質の違う HLAve スコアと PR スコアがうまく融合したためであると考えられる．HLAve 法は，表 8 が示すように，ある程度は index ページへのリンクを刈り込むことができる．また，PageRank は，トップページが最も大きなスコアとなり，トップページから遠ざかるにつれて指数的に減衰していくスコアである．よって，これらの差分をスコアとすることで，HLAve 法でリンクを刈り込んだ数に応じて index ページの重要度を下げ，PageRank の高いトップページに近いページや intensely-linked page の重要度を大幅に下げることにより，コンテンツページの重要度を強調することができる．

表 10　Web サイト@IT で HLAve-PR スコアが上位のページ

HL	HA	PR	HA-PR	URL
76	76	25	51	news/200412/17/hp.html
76	76	25	51	news/200412/17/miracle.html
76	76	25	51	news/200412/17/oracle.html
76	76	25	51	news/200412/17/storage.html
76	76	25	51	news/200412/18/cisco.html
76	76	26	50	news/200412/18/sun.html
76	76	25	51	news/200412/18/symantec.html
74	75	27	48	news/index.html
80	80	36	44	ad/adindex/index/adindex.html

5.3 考　　察

以上の実験結果から，大規模で疎な構造を持つ Web サイトに対しては，HLAve 法が効果的に index ページへの Navigate Link を刈り込み，コンテンツページに高スコアを割当てられることが分かる．

また，大規模で密な構造を持つ Web サイトに対しては，HLAve 法を単体で適用したのでは Navigate Link を効果的に刈り込めないが，PageRank スコアとの差分を取る HLAve-PR 法を適用することにより，ランキングを改善することができた．

6. ま と め

本論文では，小規模 Web サイトで有効に働くホットリンクを用いたスコアリング手法が，大規模な Web サイトに対しても同様に有効であることを，実験により検証した．これより，本手法が，サイト内検索エンジンの実践的なランキング改善手法として有効であることが示された．

今後の課題としては，より多くの Web サイトでの有効性の検証，密な構造の Web サイトに対しても効果的に Navigate Link を刈り込めるような手法の開発，HL-PR 法における正規化の方法の検討，コンテンツの解析などによる HLAve 法におけるリンクの重み付け，キーワード検索との融合などが挙げられる．また，スコアリング結果の定量的な評価方法の検討が必要である．

謝　　辞

本研究を進めるにあたり貴重なご意見をいただいた，東北大学 大学院情報科学研究科 徳山 豪 教授，本論文をまとめるにあたり有用なアドバイスをいただいた，日本アイ・ビー・エム株式会社 東京基礎研究所 松澤 裕史 氏に深く感謝申し上げます．

文　　献

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

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

[3] R. Fagin, R. Kumar, K. McCurley, J. Novak, D. Sivakumar, J. Tomlin and D. Williamson. Searching the Workplace Web. *Proc. of WWW*, pp.641-650, 2003.

[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.

[5] 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. *DBSJ Letters*, Vol.2, No.1, pp.115-118, 2003.

[6] Y. Ikawa, K. Sadakane. A Webpage Scoring Method for Local Web Search Engines. *Proc. of DASFAA*, pp.606-617, 2004.

[7] J. Chen, B. Zhou, J. Shi, H. Zhang, Q. Fengwu. Function-Based Object Model Towards Website Adaptation. *Proc. of WWW*, pp.587-595, 2001.

[8] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins. The Web as a graph: measurements, models, and methods. *Proc. of COCOON, Springer LNCS 1627*, pp.1-18, 1999.

[9] 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.

[10] T. Uno. An Algorithm for Enumerating All Directed Spanning Trees in a Directed Graph, *Proc. of ISAAC, Springer LNCS 1178*, pp.166-173, 1996.