サイト内検索エンジンのためのスコアリング手法

A Webpage Scoring Method for Local Web Search Engines

伊川 洋平† 定兼 邦彦††
Yohei IKAWA Kunihiko SADAKANE

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．

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.

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．

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．

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

† 学生会員 東北大学大学院情報科学研究科博士前期課程 ikawa@dais.is.tohoku.ac.jp
†† 正会員 九州大学大学院システム情報科学研究院情報工学部門情報基礎理論グループ sada@csce.kyushu-u.ac.jp

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．

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を基本として，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．

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．

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．

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．

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．

後の実験からも分かるように，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強調することがある．

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.

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.

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.

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.

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1,045であった. ここでは, 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1に, HotLink法によるスコアが上位のページを表2に, HL-PR法によるスコアが上位のページを表3に, スコアが下位のページを表4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.

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.

表 1 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

表 2 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

表 3 HL-PR 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
75	0	75	openwithnotepad.html
79	6	73	pinghowto.html
75	2	73	remotedesktop.html

表 4 HL-PR 法における下位のページ

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

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10,240，総リンク数は284,168であった．ここでも同様に，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5に，HotLink法によるスコアが上位のページを表6に，HL-PR法によるスコアが上位のページを表7に，スコアが下位のページを表8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．

5. まとめ

本論文では，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を用いてWebページのスコアリングを行う手法を提案した．

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．

[謝辞]

これまで本研究に関して助言してくださった皆様に深謝する．

[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.

表 5　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	62	scenter/job/index.html

表 6　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/contact_us/contact_us.html
93	38	55	aig/searchtop.html
90	44	46	applymember/club_index.html
89	28	61	scenter/learning/index.html
89	26	63	scenter/job/index.html
88	28	60	club/mail_news.html
88	17	71	news/200212/20/oracle.html

表 7　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	club/mail_news.html

表 8　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/profile.html
3	16	-13	aboutus/termofuse/termofuse.html
3	16	-13	aboutus/index.html

伊川 洋平　Yohei IKAWA

東北大学大学院情報科学研究科博士前期過程在学中．2002年東北大学工学部情報工学科卒業. Web マイニングの研究に従事．日本データベース学会学生会員.

定兼 邦彦　Kunihiko SADAKANE

2000 年東京大学大学院理学系研究科情報科学専攻終了．2000 年 4 月より東北大学大学院情報科学研究科助手．2003 年 4 月より九州大学大学院システム情報科学研究院助教授．情報処理学会，日本データベース学会正会員.