LD-2

サイト内検索システム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

1　はじめに

近年の爆発的なWebサイトの増加は，WWW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検索システムを利用しているだろう．

このWeb検索システム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のリンク構造を利用する手法に分類できる．本論文で扱うのは，後者のリンク構造を利用したスコアリングである．

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．

WWW検索システムでは，GoogleがPageRank [1] と呼ばれるスコアリング手法を実装することによって成功を収めている．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」という考え方を元に，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．

一方，サイト内検索システムでは，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，サイト内検索システムのためのスコアリング手法を提案する．

2　Webサイトのリンク構造

Webサイトのリンク構造は，トップページを根，情報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ものであると考えられる．サイト内の各ページは，単純な案内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持っていることが多い．

このようなリンク構造に対して，すべてのリンクを推薦関係としてスコアリングを行うと，トップページにもっと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．

しかし，サイト内検索を利用するユーザは多くの場合，すでにトップページにたどり着いており，葉または葉に近いレベルの情報を検索するためにシステムを利用するのである．よって，サイト内検索では，リンクの集中しやすいトップページやトップページに近いページよりも，葉または葉に近いレベルでリンクが集中しているページが重要となる．

このことから，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，サイト内検索システムにおいては好ましいとは言い難い．

一方，WWWのリンク構造は，規則性の少ない一般的なグラフであり，両者のリンク構造の間には明白な違いがある．このことからも，WWW検索システムで有効なスコアリング手法をそのままサイト内検索システムに適用するのは問題があることが分かる．

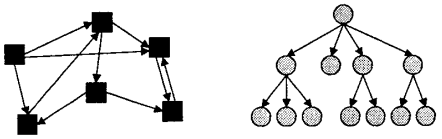

図1: WWWとWebサイトのリンク構造

3　hotlinkによるスコアリング

上記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，リンクをnavigate linkとhotlink [2] の2種類に分類する．ここで，navigate linkは単純な案内の役割を持つリンク，hotlinkは引用・推薦関係にあるリンクとして定義する．

本論文で提案するのは，Webサイト内のすべてのリンクをこの2種類に分類し，各ページがhotlinkで指されている数，すなわちhotlinkの入次数をそのページのスコアとする手法である．

ここで，Webサイト内のすべてのリンクをnavigate linkとhotlinkに選別する必要があるが，これを機械的に行う方法について考える．

Webサイトのリンク構造は，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であると考えられる．リンク構造からこの木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はあるカテゴリから別の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，本論文ではforward edgeとcross edgeをhotlink，残りのtree edgeとback edgeをnavigate link

*東北大学大学院情報科学研究科
{ikawa,sada}@dais.is.tohoku.ac.jp

として定義し，hotlink の入次数をそのページのスコアとする，hotlink によるスコアリングを提案する．

4 木の決定方法

Web サイトのリンク構造から木を決定すれば hotlink が決まり，hotlink によるスコアリングを行うことができる．ここで，Web サイトの適切な木を決定する方法が問題になる．

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 2 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．

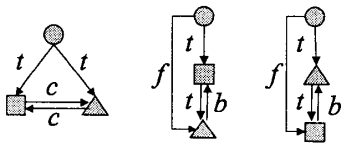

図 2: tree edge 決定の困難性

5 実験

本研究では，Windows.FAQ(http://winfaq.jp/) の Web サイトを対象データとして，PageRank と hotlink によるスコアリングの比較実験を行った．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．

それぞれの手法でスコアリングを行い，PageRank(PR) の高い順にソートした結果を表 1 に，hotlink(HL) の多い順にソートした結果を表 2 に示す．

PageRank では，トップページにスコアが集中し，続いてトップページに近いページに大きな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 によるものと考えられる．

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 によるスコアリングで上位のページである．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．

WWW 検索では強力なスコアリング手法のひとつとして知られる PageRank だが，この結果からも，そのままサイト内検索に適用するには問題があることが分かる．

一方，hotlink によるスコアリングでは，具体的なコンテンツを持ち，ある程度リンクが集中しているページに高いスコアが割り当てられている．また，トップページへのリンクはすべて hotlink にならないために，トップページのスコアは必ず 0 となる．

以上の結果から，hotlink によるスコアリングはサイト内検索システムで有効なスコアリング手法であると期待される．

表 1: PageRank の高い順にソートした結果

PR	HL	URL(http://winfaq.jp/)
100	0	index.html
30	58	w2k/index.html
24	58	wme/index.html
23	50	w98/index.html
22	50	wxp/index.html
21	25	whatsnew.html
19	83	c/9xboot.html

表 2: hotlink の多い順にソートした結果

HL	PR	URL(http://winfaq.jp/)
100	18	c/9xpref.html
100	13	c/9xdisk.html
96	16	w2k/custom.html
83	19	c/9xboot.html
79	16	w2k/boot.html
75	14	w2k/hints.html
75	13	wxp/hints.html
71	12	c/ntdisk.html

6 まとめ

本論文では，サイト内検索システム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を抽出することによって決定する hotlink の入次数をスコアとする手法を提案し，PageRank との比較実験を行った．

今後の課題としては，テキストマッチングとの連携，hotlink の重み付けやより大きな規模の Web サイトでの実験，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等が挙げられる．

参考文献

[1] 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.

[2] 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.

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