Web 検索におけるリンク構造解析
－Web サイトのグループ化と動的スコアリング－

中窪 仁† 居 行和† 佐藤 隆士‡

†大阪教育大学大学院 教育学研究科総合基礎科学専攻 〒582-8582 大阪府柏原市旭ヶ丘 4-698-1
‡大阪教育大学 情報処理センター 〒582-8582 大阪府柏原市旭ヶ丘 4-698-1
E-mail: †{nakaku, k11}@ss.osaka-kyoiku.ac.jp, ‡sato@cc.osaka-kyoiku.ac.jp

あらまし Web 上の情報を抽出する Web 検索システムの精度は未だ十分ではなく, 必要な情報に辿り着けないことも多々ある. 本研究では Web 検索結果にリンク構造解析結果を加味したスコアリングを行い, Web 検索の精度を向上させる方法を提案する. 最初に, ディレクトリ単位スコアリングおよび PageRank スコア算出コスト低減を実現する, Web サイトのグループ化手法を用いた PageRank アルゴリズムによる静的スコアリング手法について提案する. 続いて, 検索結果に応じて PageRank アルゴリズムを適用する動的スコアリング手法について提案する. 最後に, 静的・動的スコアリングを併用する手法を提案する.

キーワード 情報検索, Web とインターネット, テキスト DB, アクセスパス, ユーザインタフェース

Link Structure Analyses for Web Information Retrieval
－Grouping of Websites and Dynamic Scoring－

Hitoshi NAKAKUBO† Yukikazu KYO† and Takashi SATO‡

†Course of Mathematical and Information Science, Division of Pure and Applied Science, Graduate School of Education, Osaka Kyoiku University 4-698-1 Asahigaoka, Kashiwara, Osaka 582-8582, Japan
‡Information Processing Center, Osaka Kyoiku University 4-698-1 Asahigaoka, Kashiwara, Osaka 582-8582, Japan
E-mail: †{nakaku, k11}@ss.osaka-kyoiku.ac.jp, ‡sato@cc.osaka-kyoiku.ac.jp

Key words information retrieval, web and internet, text DB, access path, user interface

1. はじめに

Web 上の情報を抽出する Web 検索システムの精度は未だ十分ではなく, 必要な情報に辿り着けないことも多々ある. 一般的な Web 検索システムではユーザから検索語句を受け取り, その語句を含む Web ページを探し出す方法である. しかし Web ページ本文と検索語句のみに頼った手法では検索精度として限界があり, 文書構造やリンク構造解析を効果的に併用することが昨今の課題となってきている. また異なる検索方式として, 特定 Web ページに関連する Web ページを抽出するという方式も存在する. この方式については基本的にリンク構造解析を利用しているが, 検索目的と異なる Web ページ集合を抽出してしまうことも多々ある.

そこで本研究では, Web ページ本文と検索語句を利用する一般的な Web 検索システムの検索結果にリンク構造解析結果を加味したスコアリングを行い, Web 検索の精度を向上させる方法を提案する. リンク構造解析方法として, 有名な検索システム Google[1]で利用されている PageRank アルゴリズム[2]を用いることとし, 1.一定規則に則ってグループ化した Web ページ群に PageRank アルゴリズムを適用する静的スコアリング方式, 2.検索結果集合に対して PageRank アルゴリズムを適用する動的スコアリングを提案する. また, 静的スコアリングと動的スコアリングを併合する方式についても提案する.

以下, 第2章にて関連研究として PageRank アルゴリズムについて述べる. 第3章にて静的スコアリング方式に関する提案と考察を行い, 第4章にて動的スコアリング方式に関する提案と考察を行う. さらに第5章にて静的スコアリング・動的スコアリングを併合する方式について提案, 考察を行い, 第6章にてまとめる.

2. PageRank アルゴリズム

PageRank アルゴリズムは「ネットサーフィンをする人は Web 上の各 Web ページをランダムに辿る」と仮定し, 各 Web ページに辿り着く確率を元にスコアリン

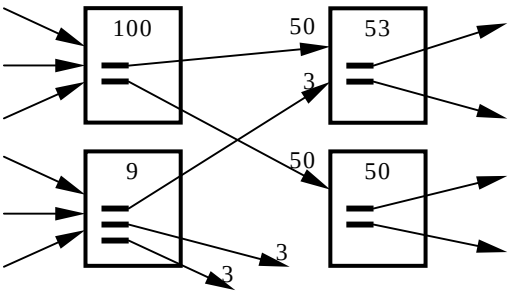

図 1 PageRank アルゴリズムイメージ
Fig. 1 Image of PageRank algorithm

グを行うアルゴリズムである. PageRank アルゴリズムのイメージを図 1 に示す. PageRank アルゴリズムにより得られるスコアは，全 Web ページに対して「被リンク数」と「リンク元 Web ページの品質」を考慮した，各 Web ページの特性を示す固定値となる．よって PageRank スコアは，Web 全体における各 Web ページの被参照度を明確に示すことになる．

しかし PageRank アルゴリズムの性質上，リンク構造において隣接関係にある Web ページ間については各 Web ページの特性が PageRank スコアに反映されやすいが，「隣接関係にはないが内容的に関連している」Web ページ間については各 Web ページの特性が PageRank スコアに反映されにくい．例えば，大きな PageRank スコアを持つ Web ページ内に外部へのリンクを持たせず，リンク構造上隣接関係にない別の Web ページに外部へのリンクをまとめてあるようなサイトでは, PageRank アルゴリズムが本来の意図通りに機能しない場合がある．

3. リンク構造解析による静的スコアリング

リンク構造上隣接関係にない Web ページ間では PageRank アルゴリズムが意図通りに機能しない場合がある問題を解決するために，総リンク数を削減してリンク構造を単純にする目的でサイトをグループ化することを考える．この場合のサイトとは「ひとまとまりに公開されている Web ページ群」であると定義し，各 Web サーバに一つ以上存在する．一般的に，サイト内には一つ以上のコンテンツが含まれており，これらのコンテンツはディレクトリ単位で管理されていることが多い．つまり，サイト内の各ディレクトリに格納されている Web ページ群は, 基本的には同分野に関するものであると考えることができる．これらをグループ化で一つの Web ページとみなすことにより総リンク数を削減する．この操作によりリンク構造上の隣接関係が拡張されるため，問題点を解決できると考えられる．

3.1. 手順

グループ化の手順として，まずリンク構造解析により一サイトとなる Web ページ群を特定する．これは，1. URL にチルダなどのユーザホームディレクトリを示すものがある場合はそのディレクトリ以下の Web ページ群, 2. URL にユーザホームディレクトリを示すものがない場合はそのドメイン以下の Web ページ群，がそれぞれ一サイトであると判断することが可能である．

次に，そのサイトのトップページ候補を選定する．トップページ候補からリンクされているサイト内 Web ページは，グループ化によりサイトトップページグループとして扱われることになる．さらにサイトトップページグループ内にディレクトリが存在している場合は，ディレクトリのトップページを選定し，ディレクトリトップページグループを形成する．

最後に，各グループからサイト外部へのリンクを全てグループトップページからサイト外部へのリンクと置き換えを行う．同様にサイト外部から各グループへのリンクについても，そのグループのトップページへのリンクとして置き換える．

この操作により，サイト内部のリンクは表面上削除され,「サイト外部からサイト内部へのリンク」は「トップページへのリンク」に,「サイト内部からサイト外部へのリンク」は「トップページから外部へのリンク」にまとめられる．

サイトのグループ化のイメージを図 2 に示す. 図内, 例 1 はグループ化の操作を行う前の状態である．これに対し，サイト内部のディレクトリについてグループ化を適用したものが例 2，サイトについてグループ化を適用したものが例 3 となる．

3.2. 考察

図 2 の各リンク構造に PageRank アルゴリズムを適用した結果を表 1 に示す．サイト外部ページである Page G，Page H に注目すると，グループ化適用範囲を大きくするに従って PageRank スコアが高くなっていることがわかる．これは，グループ化によるリンク構造上の隣接関係が拡張され，サイト外部ページへの遷移確率が高くなったためである．また，サイト内部ページである Page C，Page D，Page E に注目すると，グループ化適用範囲を大きくするに従って PageRank スコアが低くなっていることがわかる．これは，リンク構造の置き換えにより各 Web ページの本来の特性が失われてしまったことに由来すると考えられる．

これらの結果により，グループ化を行った状態で

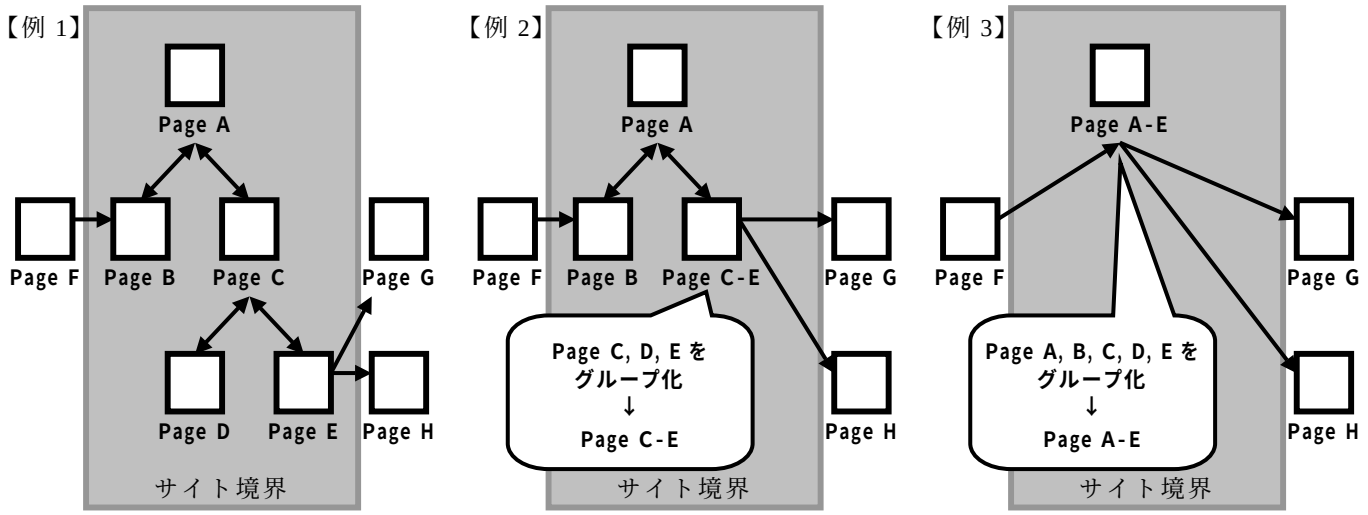

図 2 Web サイトグループ化イメージ
Fig. 2 Image of Grouping of Websites

表 1 グループ化による PageRank スコアの変化
Table 1 Change in PageRank score by Grouping of Websites

例	グループ化	サイト内部					サイト外部		
		Page A	Page B	Page C	Page D	Page E	Page F	Page G	Page H
1	なし	0.244	0.139	0.281	0.111	0.111	0.000	0.057	0.057
2	ディレクトリ	0.345	0.218	0.217			0.000	0.110	0.110
3	サイト	0.154					0.000	0.423	0.423

PageRank アルゴリズムを適用することにより, 以下のような効果が得られるといえる.

1. 「被リンク数」と「リンク元サイトの品質」を考慮した各サイトのスコアリングをすることが可能である
2. サイト内の各 Web ページがリンク構造上隣接関係か否かを意識することなくスコアリング可能となり, 前述の別の Web ページに外部へのリンクをまとめてある場合にも PageRank アルゴリズムの意図通りに機能させることが可能である
3. リンク構造の置き換えを行うことにより, 本来のリンク構造が示す各 Web ページの特性はやや失われてしまうという問題が発生する

また, PageRank スコアを算出するためには計算機上で全 Web ページのハイパーリンク構造をモデル化・数値化する必要があり, 計算コストが非常に大きいものになってしまうという問題があったが, 上記操作を行うことにより扱うべきリンク数が減少し, PageRank スコア算出時の計算コストを削減することも可能となる.

4. リンク構造解析による動的スコアリング

検索システムにおいて検索結果を出力する際, 通常は検索語出現回数を元に出力順を決定している. また Google では, 検索語出現回数に PageRank スコアを加味した上で出力順を決定することにより, 高い的中率を実現している. しかし PageRank スコアを加味する方法では, 検索結果に突出した PageRank スコアを持つ Web ページが存在しなかった場合に PageRank スコアによる重み付けの効果を失ってしまい, 通常の検索語出現回数を元にする方式とほぼ同じ出力順となってしまう可能性がある. そこで, 検索後の Web ページ集合に対してもリンク構造解析を適用することを考える.

「検索後の Web ページ集合内でのリンクは, 同分野の Web ページである可能性が高い」との考えにより, 検索後の Web ページ集合に対して PageRank アルゴリズムを適用する. この操作により高い PageRank スコアを得ることが可能な Web ページは,「同分野の記述

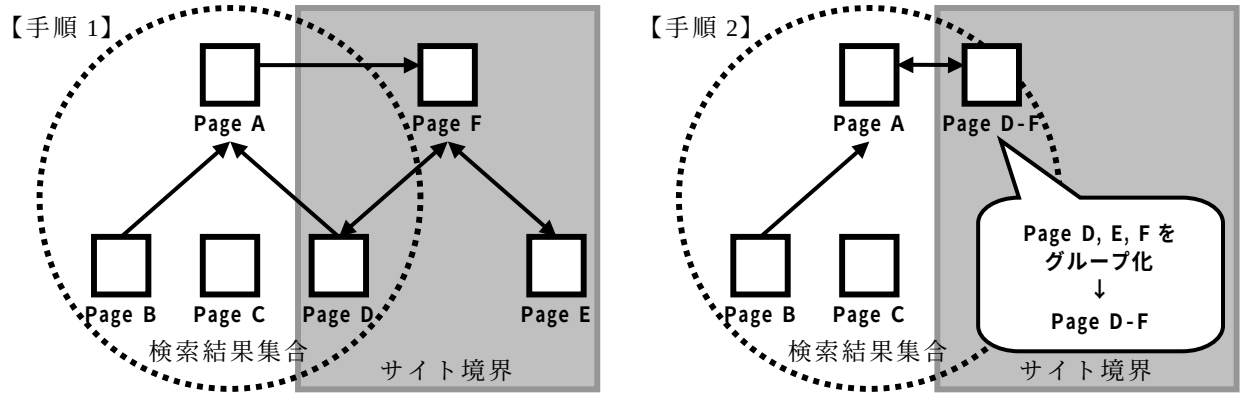

図 3 動的スコアリングイメージ
Fig. 3 Image of Dynamic Scoring

がある Web ページから多く参照されている Web ページであり，検索後の Web ページ集合において重要な位置にある」と判断することが可能である．また，これと同時に検索後の Web ページを含むサイト集合に対して，前述のグループ化を行った後に PageRank アルゴリズムを適用する．この操作により高い PageRank スコアを得ることが可能な Web サイトは，「同分野の記述を含む Web サイトから参照されている Web サイトであり，サイト内リンクを辿ることによって，検索後の Web ページ集合内の Web ページに辿り着くことができる可能性が高い」と判断することが可能である．

4.1. 手順

動的スコアリングの手順として，まず検索語句を含む Web ページ集合を抽出する．次に検索結果集合に含まれるリンク構造を抽出し，PageRank アルゴリズムを 2 回適用し，二つの PageRank スコアを算出する．1 回目に PageRank アルゴリズムを適用する時には，検索結果集合に含まれる各 Web ページのリンク構造を適用対象として PageRank スコアを算出する．2 回目に PageRank アルゴリズムを適用する時には，検索結果集合に含まれる各 Web ページを含むグループを適用対象にする．

この操作により，1 回目に得られる PageRank スコアは検索結果集合内の各 Web ページ間のリンク構造および各 Web ページの特性を生かしたスコアリング結果となる．また 2 回目に得られるスコアは，リンク構造上隣接関係にない検索結果集合内の各 Web ページ間の関係を考慮したスコアリング結果となる．この二つの数値を合わせることにより，動的に精度の高いスコアリングを行うことができると考えられる．ただし，2 回算出した PageRank スコアは同等の重みを与えることができるものではない．1 回目で得られた PageRank スコアの方が 2 回目で得られるスコアよりもより厳密なリンク構造解析を行った結果となると思われるので，2 回目で得られた PageRank スコアに定数 n（$0<n<1$）を乗算するなどの重み付けが必要となる．

以上の手順のイメージを図 3 に示す．図内，手順 1 が 1 回目の PageRank アルゴリズム適用，手順 2 が 2 回目の PageRank アルゴリズム適用を示す．

4.2. 考察

図 3 のリンク構造に PageRank アルゴリズムを適用した結果を表 2 に示す．なお，1 回目と 2 回目の PageRank スコアを併合したスコアを手順 3 として記載している．また手順 3 での重みは，定数 $n = 0.5$ として算出している．

手順 1 では Page A のみが検索結果集合内にてリンクされている Web ページであるため，Page A のスコアが突出する結果になる．しかし手順 2 では Page D を含むグループである Page D-F が Page A よりリンクされているため，Page D にも PageRank スコアが加算されていることがわかる．よって手順 3 での併合の結果，「Page A ＞ Page D ＞ Page B ＝ Page C」となる．

以上より，Page A は「同分野の記述がある Web ページから多く参照されている Web ページであり，検索後の Web ページ集合において重要な位置にある」と判断されていることがわかる．また Page D は，「同分野の記述を含む Web サイトから参照されている Web サイトであり，サイト内リンクを辿ることによって，検索後の Web ページ集合内の Web ページに辿り着くことができる可能性が高い」と判断されていることがわかる．

ゆえに，動的スコアリング法を用いることにより，以下のような効果が得られると言える．

表 2 動的スコアリングで得られる PageRank スコア
Table. 2 PageRank score of Dynamic Scoring

手順	Page Rankアルゴリズム 適用範囲	検索結果集合内				検索結果集合外	
		Page A	Page B	Page C	Page D	Page E	Page F
1	検索結果集合内	1.000	0.000	0.000	0.000	-	-
2	検索結果集合に一部でも含まれるグループ	0.504	0.000	0.000	0.496		
3	-	1.252	0.000	0.000	0.248	-	-

※(手順 3 スコア)＝(手順 1 スコア)+{(手順 2 スコア)× 0.5}

1. 検索結果集合内での被参照度を元にしたスコアリングを行うことが可能である
2. 検索結果集合内の Web ページが含まれる Web サイトにもスコアリングを行うことが可能である．その結果，検索結果集合内でリンク構造上の隣接関係にない Web ページにも適切なスコアリングを行うことが可能である．

また，動的スコアリング法で得られる値は検索語により変動する値であるため，静的スコアリングに比べてより検索語に特化したスコアリングを行うことが可能であるといえる．

ただし，1 回の検索要求につき 1 回のリンク構造抽出処理と 2 回の PageRank スコア算出処理が必要となるために，時間計算量が大きくなってしまう問題がある．また 2 回算出する PageRank スコアの重み付けについて，定数 *n* の最適値を検討する必要がある．

5. 静的・動的スコアリングの併合

本研究の静的アルゴリズムと通常の PageRank アルゴリズムとを比較した場合，静的スコアリングは Web サイトのグループ化により発生する各 Web ページ特性の損失があるため，PageRank アルゴリズムに比べると精度面では劣る．しかし，時間計算量の面では本研究の静的スコアリングの方がリンク構造の削減を行う分優れていると考えられる．また，動的スコアリングと PageRank アルゴリズムとを比較した場合は，本研究の動的スコアリングの方が精度面では優れ，時間計算量の面では劣ると考えられる．

しかし本研究で提案した 2 つのスコアリング法は，静的スコアリングで問題となる精度面を動的スコアリングで補填し，動的スコアリングで問題となる時間計算量面を静的スコアリングで補填することが可能である．

そこで，静的スコアと動的スコアを併合したスコアを算出することを考えると，Web サイトのグループ化を行うことにより発生する各 Web ページ特性の損失を動的スコアリングによって補填でき，さらに動的なスコアリングを行うため，最終的には PageRank アルゴリズムより良い結果が得られると考えられる．また Web サイトのグループ化により静的スコアリングのコストが抑えられるため，Web ページデータベースの更新頻度によっては PageRank アルゴリズムよりも低コストで運用することも可能と思われる．

また別のアプローチによるスコアリング手法である HITS アルゴリズム[3]と比較した場合，Web サイトのグループ化による各 Web ページ特性の損失，特にリンク構造で隣接する玉石混合の Web ページ群が同程度にスコアリングされてしまうという点は，HITS の常に的確なコミュニティを抽出できるわけではない問題点と類似している．しかしこの問題は，動的スコアリングにより解消されるものであり，結果，精度面では HITS アルゴリズムより優れていると考えられる．

6. おわりに

今回は，リンク構造解析を利用した Web 検索精度向上方法として．各 Web サイトの特性を利用した静的スコアリング，各 Web サイトおよび各 Web ページの特性を利用した動的スコアリング，静的スコアリング，動的スコアリングを併合したスコアリングの手法を提案した．これらのスコアと検索語出現回数を加味したスコアリングを行うことにより，各 Web ページの特性だけでなく各 Web サイトの特性を考慮したスコアリングを行うことが可能となり，精度向上を望むことが可能となると考えられる．

今後は，今回提案したスコアリング法を利用して実際どの程度の精度向上が見られるかについて検証していく．その過程で，1.静的スコアリングでの適切なグループ化範囲，2.動的スコアリングで 2 回算出する PageRank スコアの適切な重み，3.併合スコアリングでの静的スコア，動的スコアの適切な重み，のそれぞれについて検討を行っていく．また，動的スコアリングは結果出力までに時間を要することが明白であるため，速度向上についても検討していく予定である．

文　献

[1] S. Brin and L. Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, In Proc. of the 7th International World Wide Web Conference (WWW7), pp.107-117, 1998.

[2] L. Page, The PageRank Citation Ranking: Bringing Order to the Web, http://google.stanford.edu/~backrub/pageranksub.ps, 1998.

[3] J. M. Kleinberg, Authoritative Sources in a Hyperlinked Environment, Journal of the ACM, vol.46, no.5, pp.604-632, 1999.

[4] 小畑喜平，リンク情報を利用した WEB 検索システム，平成 14 年度大阪教育大学大学院修士論文.