DEWS2006 3B-i4

文書集合単位のリンク解析を用いた Web ページスコアリング

中窪　仁† 中島　伸介† 波多野賢治†
宮崎　純† 植村　俊亮†

† 奈良先端科学技術大学院大学 情報科学研究科
〒 630-0192 奈良県生駒市高山町 8916-5
E-mail: †{hitosh-n,shin,hatano,miyazaki,uemura}@is.naist.jp

あらまし 本稿では，Web ページ単位のリンク解析ではなく，Web ページ集合単位でのリンク解析を用いた Web ページスコアリング手法を提案する．PageRank 等の既存のリンク解析スコアリング手法では，Web ページの重要度の評価をページ単位で行っている．しかし，一人の著者が一つの話題に関して作成した Web コンテンツは，単一の Web ページのみに存在することは少なく，複数の Web ページにまたがって存在することが多い．そこで，同一の著者が記述し，同一の話題を扱う Web ページ集合を一つの文書集合 WPS として扱い，その文書集合間のリンク構造を解析することにより，Web ページ集合単位で重要度の評価を行う．NTCIR テストコレクションを利用して，提案手法を既存手法と比較実験したところ，二種類の検索精度評価指標において提案手法の優位性を確認することができた．
キーワード Web 情報検索，WPS，リンク解析，NTCIR

Web Page Scoring based on a Link Analysis of Web Page Sets with Same Contents

Hitoshi NAKAKUBO†, Shinsuke NAKAJIMA†, Kenji HATANO†,
Jun MIYAZAKI†, and Shunsuke UEMURA†

† Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0192, Japan
E-mail: †{hitosh-n,shin,hatano,miyazaki,uemura}@is.naist.jp

Abstract We propose a Web page scoring method based on a link analysis among sets of Web pages. Conventional link analysis methods like PageRank and HITS calculate importance degree of each Web page. However, authors often create multiple Web pages to describe a specific topic. The importance degrees of such multiple Web pages can not be calculated by the conventional link analyses accurately. To cope with this problem, we have to treat Web pages with same contents which an author edited as one Web page set (WPS). After constructing link structure among WPSs, we calculate importance degrees of WPSs using conventional link analysis methods and then, importance degrees of Web pages based on that of WPSs. In this paper, we compared our approach with conventional ones using NTCIR test collection, and found our approach is better than them in WRR and DCG evaluations.
Key words Web information retrieval, WPS, Link analysis, NTCIR

1. はじめに

Web 検索エンジンは，Web 空間上に存在する膨大な情報の中から必要な情報を探し出すための手段として利用されている．しかし，現在の Web 検索エンジンではユーザの検索要求を満たすことができない場合もある．例えば，検索語句は含まれているがユーザの要求する情報が含まれていない Web ページが検索結果の上位に出てきてしまう場合である．そこで，このような問題を解決する，より多くのユーザの検索要求を満たすことができる Web 検索エンジンが必要であると考えられる．

Web 情報検索における検索精度向上に有効な手法として，Web ページ間のリンク構造を利用する手法が知られている．特

に PageRank アルゴリズム [1] と HITS アルゴリズム [2] は，リンク解析を利用する代表的な手法である．PageRank アルゴリズムはランダムウォークモデルをシミュレートし，リンク構造上の重要な Web ページに加点することを再帰的に繰り返し，各 Web ページの重要度を算出する．また，HITS アルゴリズムは Authority と Hub の二つの概念を利用して各 Web ページの重要度を算出する．

これらの手法は，リンク構造の解析や評価を Web ページ，すなわち一つの HTML ファイル単位で行っている．しかし，一人の著者が一つの話題に関して作成した Web コンテンツは単一の Web ページに存在することは少なく，Web サイト内の複数の Web ページに分散して存在することが多い．また Web サイトによっては，複数の Web ページを順に閲覧することを前提に作成されたものもある．このように複数の Web ページから一つの内容を表現しているような場合は，ページ単位でリンク解析を行い各 Web ページの重要度を計算するような既存手法では，Web ページの重要度を正しく計算できない可能性がある．

そこで我々は，Web ページ単位のリンク解析ではなく，Web ページ集合単位 (Web コンテンツ単位) でリンク解析を行い，その結果を元に Web ページの重要度を計算する手法を提案する．提案手法を適用することで，Web コンテンツに基づいた重要度の計算が可能となるため，より正確に各 Web ページの重要度計算ができるようになる．

以下 2. で関連研究について述べる．続いて 3. で本稿で提案手法について詳述し，4. で本提案の評価実験を行う．そして 5. で，まとめと今後の課題について述べる．

2. 関 連 研 究

複数の Web ページの情報を用いて，Web 情報検索における検索精度の向上に役立てようとした研究は，過去にも多く存在する．

杉山らは，ある Web ページの特徴ベクトルを，その Web ページからリンクされている複数の Web ページをクラスタリングした上で，そのクラスタ重心ベクトルを用いて修正することで，検索精度の向上を図っている [3]．また正田らは，同一 Web サイト内に存在する複数 Web ページに対してクラスタリングを行い，その結果を元に Web ページの文書ベクトルを修正して，検索制度の向上を図っている [4]．これらの研究は，Web ページの内容を特徴ベクトルで正しく表現するために，複数の Web ページの情報を用いたものである．

一方 Tajima らは，利用者が入力した問合せキーワード全てを含む最小部分グラフを Web 空間を表現するグラフ構造から抽出し，それらのスコアを計算することで検索結果を提示する Web 検索エンジンを開発している [5]．また，同様の研究として Li らによる “information unit” の概念もある [6]．これらの研究に共通なのは，検索結果が Web ページではなく，リンクで結びついた複数の Web ページであるという点である．

以上で挙げた既存手法は，Web のリンク構造を Web ページの特徴ベクトルの修正や検索単位の変更に利用することで，最終的に Web 情報検索における検索精度の向上を図っている．しかし，PageRank や HITS などのリンク解析による Web 情報検索が主流になってきた現在では，リンク解析の手法を取り入れつつ，1. に挙げた問題点を解決する方法を考えるのが自然である．

本稿で我々が提案する手法は，既存のリンク解析による Web ページの重要度計算を，Web ページ集合単位で行うことにより精度向上を図る手法であり，既存のリンク解析の手法を取り入れつつ，上記研究の目的を実現しようとしたものである．

3. 文書集合単位のリンク解析を用いた Web ページスコアリング

本稿で取り上げた問題点は，複数の Web ページから一つの内容を表現しているような場合は，既存のリンク解析による Web ページスコアリング手法は Web ページの重要度を正しく計算できない可能性があるという点である．したがって，既存のリンク解析を適用する前に Web ページから同一の内容を表現している Web ページ集合を抽出する必要がある．その後，実際の Web ページ間のリンク構造から Web ページ集合間のリンク構造を構築し，その構造を利用してリンク解析することにより Web ページ集合単位での重要度計算を行う．

3.1 文書集合の定義

本稿では，リンク解析の単位とする文書集合を「同一の著者が記述し，同一の話題を扱う Web ページ集合」と定義する．

これは，以下の経験則による．

- 同一の著者が記述

Web ページは，著者によって扱う話題やその性質が異なると考えられる．例えば，図 1 のように，同じ「データベース」の話題を扱った Web ページであっても，著者がデータベース学の権威であった場合，その内容は比較的学術的で技術レベルの高い内容になると考えられる．一方，著者がデータベース初心者であった場合，その内容は SQL の使い方の初歩などの入門レベルのものになると考えられる．このような内容や質が異なる情報を同一の重要度とすることは，それぞれの情報の過小評価や過大評価につながるため，正確な重要度を決定できなくなると考えられる．よって本手法において解析単位となる文書集合は，「同一の著者が記述した Web ページ集合で構成されたもの」と定義する．

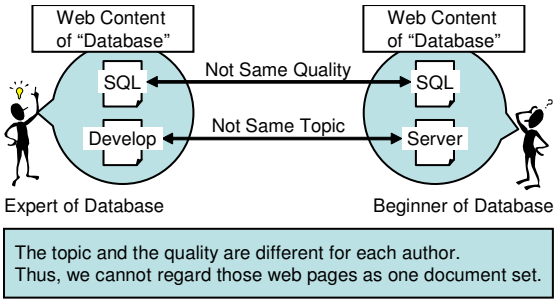

図 1 文書集合作成の方針: 同一の著者
Fig. 1 Policy of Document Set Making: Same Author

なお，同一著者によって作成された Web ページ集合の分割

には，Web サイトの境界を利用する．これは，Web サイトの管理者が Web サイト内の Web ページを作成している，という仮定に基づく．ここで Web サイトとは，「一定の内容，ルールおよびデザインをもつ Web ページ群」であり，入り口となる Web ページ (エントリページ) を一つだけ持つものとする．

- 同一の話題を扱う

同一著者によって作成された Web ページ集合であっても，著者の特性によって情報の性質が異なると考えられる．例えば図 2 のように，一人の著者が「データベース学」「スポーツ科学」の二分野に関する Web ページを作成していたとする．このとき，この著者の「データベース学」と「スポーツ科学」に関する理解度は，同じレベルであるとは限らないため，Web ページの重要度を計算する場合には，別の重要度が計算されるべきであると考えられる．よって，本提案において解析単位となる文書集合は，「同一分野に含まれる Web ページ集合で構成されたもの」と定義する．

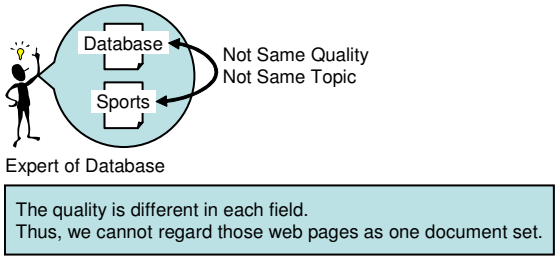

図 2　文書集合作成の方針: 同一の話題
Fig. 2　Policy of Document Set Making: Same Topic

本稿では，この概念から決定された Web ページ集合を WPS (Web Page Set) と呼ぶことにする．

3.2　処 理 手 順

本提案手法の処理手順は図 3 のとおりである．以下にその内容を詳述する．

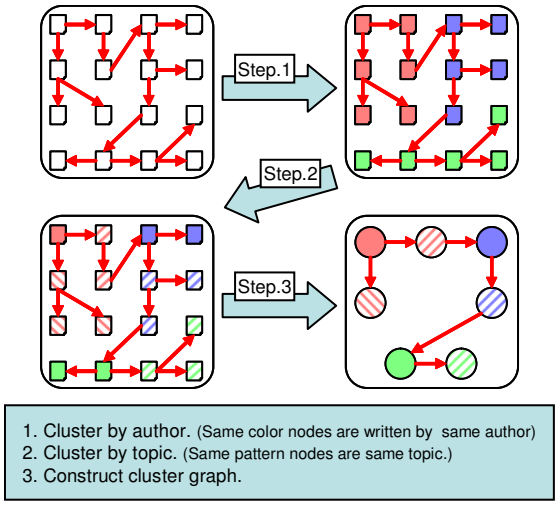

図 3　提案手法の手順
Fig. 3　Procedure of Our Proposal Method

（1）　Web サイトの境界決定

Web サイトの境界決定は，Ayan らの手法 [7] を利用して行う．これは物理ドメイン (Web サーバ) から論理ドメイン (Web サイト) を抽出する手法であり，エントリページ候補決定，論理ドメイン境界決定，という二つの操作で構成される．

エントリページ候補決定:　URL 文字列，リンク数が一定条件を満たした場合にエントリページ候補スコアを各 Web ページに加算していく．最終的なスコアが高い Web ページから順にエントリページ候補として扱う．

論理ドメイン境界決定:　物理ドメイン上の Web ページをディレクトリツリーとして扱い，エントリページ候補自身とエントリページ候補の子以下を同一論理ドメインとする．論理ドメインを構成する Web ページ数が閾値以下の場合は，上位の論理ドメインに併合を行う．本稿では，Web ページ数の閾値を実験的に 10 とした．

（2）　分野の境界決定

分野の境界決定には，Web ページの内容に基づくクラスタリング手法を用いる．Web ページの特徴ベクトルは，ChaSen [8] による形態素解析と，TF-IDF 法による重み付けを用いて決定する．またクラスタリングには，分類感度が高いとされるウォード法 [9] を利用する．生成クラスタ数は，1 クラスタにつき約 10 Web ページが割り当てられるクラスタ数を考え，実験的にクラスタリング対象数の 1/10 とした．

（3）　WPS 間のリンク構造の構築

(1) と (2) の条件を満足する Web ページ集合を一つの WPS と決定する．本稿ではこうして決定した WPS を一つのコンテンツとして扱う．WPS 間のリンク構造は，以下の操作を Web ページ間のリンク構造に適用することで構築する．

（a）　同一 WPS 内 Web ページ間のリンクを削除する．

（b）　リンクの始点終点をそれぞれ，その Web ページが属する WPS に変更する．

（4）　PageRank による WPS の重要度計算

WPS 間に存在するリンク構造を利用して，各 WPS の重要度を計算する．最終的に，WPS に含まれる Web ページに対して重要度の計算を行う必要があるが，今回は実験的に WPS の重要度をそのまま Web ページに対して付与する．

4.　評 価 実 験

本節では提案手法の有効性を判断するための評価実験を行い，既存手法との比較を行う．

4.1　実 験 環 境

実験対象には，NTCIR-4 Web [10], [11] で使用されたテストコレクションである NW100G-01 を利用した．これは，Web ページ総数が約 2,370 万 Web ページ，リンク総数が約 8,000 万リンクのデータである．また，検索課題には，NTCIR-4 Web Task A (情報指向検索) で使用された検索課題を利用した．実験結果の評価は，NTCIR-4 Web Task A にて使用された適合判定結果を元に行った．これは，多値適合レベルによって四段階に判定されたもので，評価尺度には，NTCIR-4 Web で利用された Weighted Reciprocal Rank (WRR) [12], [13] および Discounted Cumulative Gain (DCG) [14] を利用している．

WRR は，主に初出の適合文書がどの程度上位に現れるかを評価する尺度であり，Mean Reciprocal Rank (MRR) [15] を多値適合レベルに対応するように拡張した評価手法である．MRR は，しばしば質問応答システムの評価に利用される評価方式であり，その値は各質問に対する結果リストにおける初出回答のランクの逆数を，全質問にわたって平均した値である．一方 DCG は多値適合レベルに適した評価尺度であり，適合文書のランクを考慮することにより適合度順の評価を行う．その値は，各ランクにおける適合値を加算した値である．

なお，本実験で用いた Web 検索エンジンには，可変長グラムベースインデクスを利用した全文検索システム [16] を利用した．

4.2 結　　果

図 4 と図 5 は，PageRank アルゴリズムを用いた場合の Web ページランキングと，本稿で提案した WPS に対し PageRank アルゴリズムを適用した Web ページランキングを，WRR および DCG を元にプロットしたグラフを表している．両グラフにおいて，本稿の提案手法 (Our Method) が，PageRank アルゴリズムによる手法 (Conventional) よりも上位にプロットされていることが容易に見て取れる．このグラフを細かく分析した場合，図 4 より提案手法が初出の適合文書をより上位にランキングできていることが，また，図 5 より提案手法が既存手法より多くの適合文書を上位に抽出できていることが判明した．つまりこれらの結果は，本提案が既存手法より検索精度の点で優位性を持っていることを示している．

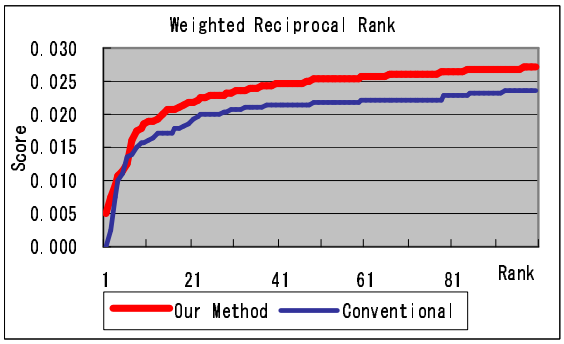

図 4　評価結果 (WRR)
Fig. 4　Evaluation Result (WRR)

以上のことから，PageRank などの Web のリンク構造解析は，Web ページ単独で行うよりも，Web ページのコンテンツに基づいて行うほうが，Web ページの重要度を正しく計算することができることが判明した．

5. おわりに

本稿では，Web ページ単位のリンク解析ではなく，Web ページのコンテンツ単位 (WPS) でのリンク解析を用いた Web ページスコアリング手法を提案した．また，既存手法と提案手法を Web 情報検索のテストコレクションの一つである NTCIR の二種類の検索精度評価基準を用いて評価し，提案手法が既存手法よりも適合文書を上位に検索可能であることを確認した．

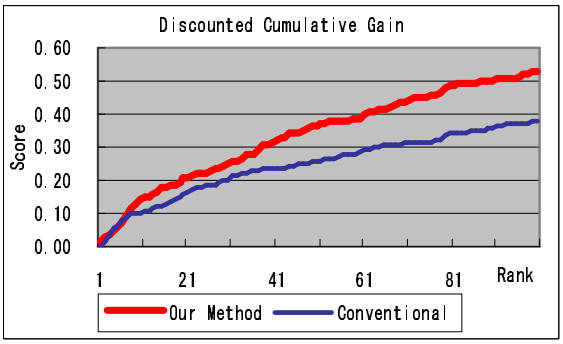

図 5　評価結果 (DCG)
Fig. 5　Evaluation Result (DCG)

しかし，提案手法は WPS 単位でスコアリングを行い，そのスコアをそのまま WPS 内の Web ページに付与するため，検索精度が WPS の決定方法に依存することになる．また，同じ WPS 内に含まれている Web ページに同じスコアが与えられるため，これらの問題点を解決するための提案，すなわち現在のクラスタリング手法，生成クラスタ数の決定法や，WPS の決定方法の妥当性の確認，同一 WPS 内の Web ページに対するスコアリング法の提案を行う必要がある．

謝辞　本研究の一部は科研費 (基盤研究 (A) (2) 課題番号: 15200010，若手研究 (B) 課題番号: 17700132) により行われた．ここに記して謝意を表す．

文　　献

[1] S. Brin and L. Page: “The anatomy of a large-scale hypertextual web search engine”, Proceedings of the 7th World-Wide Web Conference (WWW7) (1998).
[2] J. Kleinberg: “Authoritative sources in a hyperlinked environment”, ACM-SIAM Symposium on Discrete Algorithms (1998).
[3] 杉山, 波多野, 吉川, 植村：“ハイパーリンクで結ばれた隣接ページの内容に基づく web ページのための tf-idf 法の改良”, 電子情報通信学会論文誌, **J87-D-I**, 2, pp. 113–125 (2004).
[4] T. Masada, A. Takasu and J. Adachi: “Link-based clustering for finding subrelevant web pages”, Proceedings of the Third International Workshop on Web Document Analysis (WDA2005) (2005).
[5] K. Tajima, K. Hatano, T. Matsukura, R. Sano and K. Tanaka: “Discovery and retrieval of logical information units in web”, Proceedings of the Workshop on Organizing Wep Space (WOWS 99), pp. 13–23 (1999).
[6] W.-S. Li, K. S. Candan, Q. Vu and D. Agrawal: “Retrieving and organizing web pages by “ information unit ””, WWW ’01: Proceedings of the 10th international conference on World Wide Web, New York, NY, USA, ACM Press, pp. 230–244 (2001).
[7] N. F. Ayan, W.-S. Li and O. Kolak: “Automating extraction of logical domains in a web site”, Data Knowl. Eng., **43**, 2, pp. 179–205 (2002).
[8] “ChaSen”, http://chasen.naist.jp/.
[9] 神嶌：“データマイニング分野のクラスタリング手法 (1) －クラスタリングを使ってみよう！ －”, 人工知能学会誌, **18**, 1, pp. 59–65 (2003).
[10] K. Eguchi, K. Oyama, A. Aizawa and H. Ishikawa: “Overview of web task at the fourth ntcir workshop”, Working Notes of the Fourth NTCIR Workshop Meeting (2004).
[11] K. Eguchi, K. Oyama, A. Aizawa and H. Ishikawa: “Overview of the information retrieval task at ntcir-4 web”,

Working Notes of the Fourth NTCIR Workshop Meeting (2004).

[12] K. Eguchi, K. Oyama, E. Ishida, N. Kando and K. Kuriyama: "Overview of the web retrieval task at the third ntcir workshop", Technical Report NII-2003-002E, NII (2003).

[13] K. Eguchi, K. Oyama, E. Ishida, N. Kando and K. Kuriyama: "Evaluation methods for web retrieval tasks considering hyperlink structure", IEICE Transactions on Information and Systems, **E86-D**, 9, pp. 1804–1813 (2003).

[14] K. Järvelin and J. Kekäläinen: "Ir evaluation methods for retrieving highly relevant documents", Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 41–48 (2000).

[15] E. Voorhees: "The trec-8 question answering track report", Proceedings of TREC-8, pp. 77–82 (1999).

[16] T. Sato, T. Satomoto and K. Han: "Ntcir-3 pat experiments at osaka kyoiku university", Working Notes of the Third NTCIR Workshop Meeting Part III: Patent Retrieval Task, Tokyo, Japan, pp. 21–24 (2002).