Web ページのグループ化による静的動的ランキング

中窪 仁† 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@ss.osaka-kyoiku.ac.jp, ‡sato@cc.osaka-kyoiku.ac.jp

あらまし インターネットの普及に伴い，WWW 空間上に蓄積される情報が急増している．この情報群から検索を行うために Web 検索システムが存在するが，精度的に十分とはいえない．そこで我々は Web 検索システムの精度向上を図るため，Web ページのグループ化とリンク構造解析を併用した手法を提案する．この手法は，リンク構造上隣接関係の拡張を行うための Web ページグループ化手法，グループ化を利用したリンク構造解析による静的スコアリング手法，検索語句に依存する検索結果集合のリンク構造解析にグループ化を加味した動的スコアリング手法，および静的スコアと動的スコアを併用するランキング手法から構成される．本論文ではこれらの手法について述べ，実験結果を報告する．

キーワード 情報検索，Web マイニング，インデックス

Static and Dynamic Ranking by Web Page Grouping

Hitoshi NAKAKUBO† 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@ss.osaka-kyoiku.ac.jp, ‡sato@cc.osaka-kyoiku.ac.jp

Keyword Information Retrieval, Web Mining, Index

1. はじめに

インターネットの普及に伴い，WWW 空間上に蓄積される情報量は増加しつづけている．この情報群は非常に大規模であり，情報の質も様々であるため，インターネット利用者が必要とする情報を抽出する事は難しい．この困難な作業を補助するシステムとして，Web 検索システムが存在する．しかし，Web 検索システムを利用したとしても常に必要な情報を抽出できるとは限らず，Web 検索システムのさらなる検索精度向上が望まれている．

現在利用されている一般的な Web 検索システムは，インターネット利用者が入力した検索語句を基に各 Web ページ本文を検索することにより適合 Web ページ群を抽出するものである．しかし検索語句と Web ページ本文のみを利用した，単純な全文検索手法により得ることのできる検索精度には限界がある．そこで，Web ページ特有の情報である HTML 文書構造やリンク構造を効果的に利用した手法が検討されている．

リンク構造解析を利用した代表的な手法として，有名な検索エンジンである Google[1]にて採用されている手法である PageRank アルゴリズム[2][3]，およびコミュニティ抽出にも効果的な手法である HITS アルゴリズム[4]がある．これらは各 Web ページ間の隣接関係を基にランキングを決定する手法であり，隣接関係にない Web ページとの関係は再帰的に解決される．

しかし，現実の WWW 空間上では関連する Web ページ間で常にリンク構造が存在するとは限らない．例えば，リンク行為を Web サイトトップページにしか許可していない Web ページが存在した場合，隣接関係を基に決定したランキングでは適切な結果を得ることができないと考えられる．

我々はこの問題を軽減するため，類似情報を持つ Web ページ群をグループ化することにより，リンク構造上の隣接関係を拡張する手法を提案する．また，リンク構造隣接関係の拡張による検索精度の低下を防止するため，動的スコアリング手法を提案する．

以降，第 2 章にて関連研究およびそれらの持つ問題点について述べる．続いて第 3 章にて提案手法について述べる．第 4 章にて実験環境について，第 5 章にて実験結果について示す．第 6 章にて考察を行い，第 7 章にてまとめる．

2. 関連研究

2.1. PageRank アルゴリズム

PageRank アルゴリズムは，1998 年にスタンフォード大学の L. Page と S. Brin が提案した Web ページの重要度を表す指標である．この手法はリンク行為を「リンク先 Web ページの推薦行為」と捉え，リンク先 Web ページにスコアを分け与えるアルゴリズムとなってい

る．この手法により得られるスコアは各 Web ページの被参照度を明確に示す値となる．しかし，推薦を行いたい Web ページに対して常にリンク可能であるとは限らないという点が問題になると考えられる．

2.2. HITS アルゴリズム

HITS アルゴリズムは，1997 年に IBM の J. M. Kleinberg により提案されたコミュニティ抽出のための手法である．これは Web ページの重要度を表す指標として Authority と Hub を定義し，この二つの関係を「よい Authority は複数の良質の Hub によってリンクされ，また良質の Hub は複数のよい Authority にリンクをしている」と定義している．この手法により得られるスコアは，Authority は情報源として，Hub はリンク集として有用な Web ページ群を抽出可能な値となる．しかし，コミュニティ候補抽出に全文検索結果を利用していないため，常に適切なコミュニティを抽出できるとは限らないという問題点を持つ．

3. 提案手法

我々の提案する手法は，グループ化，静的スコアリング，動的スコアリング，およびランキングからなる．以下にてそれぞれについて簡単に述べる．詳細は文献[5]を参照のこと．

3.1. グループ化

リンク構造上隣接関係を拡張するため，類似情報を持つ Web ページ群をグループ化することを考える．我々は類似情報を持つ Web ページ群を「同一作成者により作成された，類似情報を持つであろう Web ページ群」と定義した．グループ化処理にはディレクトリ構造を利用することとした．この場合，「同一作成者」は Web サイト区切りを，「類似情報を持つであろう Web ページ群」は同ディレクトリ内に含まれる Web ページ群を，それぞれ URL 文字列より決定することによりグループ化を行うことが可能である．

リンク構造上隣接関係の拡張は，グループ内 Web ページ間でのリンク構造を削除し，グループ間のリンク構造に置換することにより実現される．

3.2. 静的スコアリング

静的スコアリングは，WWW 空間上の全 Web ページにグループ化を適用した後のリンク構造に対してリンク構造解析スコアリングを行う．グループ化を行う事により，PageRank アルゴリズムにおける問題点を軽減できると考えられる．

3.3. 動的スコアリング

動的スコアリングは，全文検索結果集合に含まれるリンク構造を用いてリンク構造解析スコアリングを行う．その際，全文検索結果集合内の Web ページ間で構成されるリンク構造を対象にしたリンク構造解析スコアと，全文検索結果集合内の Web ページをグループとして扱った状態で構成されるリンク構造を対象にしたリンク構造解析スコアの二つを算出する．これらはそれぞれ，明確な被参照度を示すスコア，リンク構造上隣接関係を拡張したスコア，と特徴づけることが可能である．全文検索結果集合内の Web ページを基として動的スコアを算出するため，HITS アルゴリズムにおける問題点は回避できると考えられる．

3.4. ランキング

全文検索スコアに前述の手法にて得られた静的，動的スコアを併合することにより併合スコアを算出し，それを基にランキングを行う．併合スコアは各スコアを正規化し，それぞれに重み係数を乗算した後に加算することで算出する．

4. 実験環境

実験に使用した環境は以下のとおりである．

全文検索システム：可変長グラムベースインデクス[6]を使用した．全文検索結果抽出数は上位 2,500 件とし，提案手法適用後の評価には上位 1,000 件を用いた．

検索対象：NTCIR-4 Web Task[7][8]にて提供されたテストコレクション NW100G-01 を使用した．これには Web ページ総数 1,100 万ページ，リンク総数 8,000 万リンクが含まれる．

検索課題：NTCIR-4 Web Task B にて提供された検索課題のうち，NTCIR-4 Web Task B で実際には使用されなかった課題，および全文検索結果抽出数が 2,500 件に満たなかった課題を除いた 77 課題を利用した．

評価方式：NTCIR-4 Web Task にて採用されている WRR[9][10]，および 11 点平均適合率を利用した．

5. 実験結果

5.1. グループ化および静的スコアリング

検索対象にグループ化を適用した結果を表 1に示す．各グループに含まれる Web ページ数にばらつきがあることがわかる．Web ページ数のばらつきはリンク数に影響すると考えられ，グループ化手法を再検討する必要があると考えられる．

グループ化適用前後による比較結果を表 2に示す．スコア最大値はグループ化前の方が高いがスコア最小値はグループ化後の方が高いことがわかる．これは，ある適合文書についてグループ化前スコアの適用が効果的な場合もあれば，グループ化後スコアの適用が効果的な場合もあるということである．また，スコア範囲が大きく異なることがわかる．これは，グループ化前スコアはランクに与える影響が大きく，グループ化後スコアはランクに与える影響が小さいことがわかる．

グループ化前後それぞれについて，適合文書を抽出した検索課題の割合を図 1に示す．図中，「双方」はグループ化前後両方で抽出できた課題を表し，補助グラフにてグループ化前後どちらがより良い評価を得られたかを表す．抽出可能な検索課題の割合は，グループ化前については 61%，グループ化後については 13%であることがわかる．これは，グループ化後で抽出可能な検索課題はグループ化前に比べて少ないが，グループ化前では抽出不可能な検索課題を 12%抽出可能であることを意味している．

これらの結果より，グループ化前後それぞれのスコアは互いに逆の性質を持っていると考えられる．ゆえに二つのスコアを併合することで両方の性質を併せ持つスコアリングをすることが可能であると考えられる．静的スコア *ScoreS(p)*の算出式を以下に示す．なお，*Retrieval(p)*，*StaticN(p)*，*StaticG(p)*はそれぞれ文書*(p)*における全文検索スコア，グループ化前静的スコア，

グループ化後静的スコアを，*Wr*，*Wsn*，*Wsg* は各スコアの重み係数を表す．

$$ScoreS(p) = Wr \cdot Retrieval(p) + Wsn \cdot StaticN(p) + Wsg \cdot StaticG(p)$$

図 2に静的スコアリング評価結果を示す．結果より，(*Wr*，*Wsn*，*Wsg*)=(1,1,1)にて精度向上を確認できた．

5.2. 動的スコアリングおよびランキング

全文検索結果集合にグループ化を適用した結果を表 3に示す．スコア最大値最小値の関係は静的スコアリング時と同様だが，グループ化後のスコア中央値がスコア平均値以上となり，分布自体はグループ化前後で似た傾向となった．

グループ化前後それぞれについて，適合文書を抽出した検索課題の割合を図 3に示す．図中，「双方」はグループ化前後両方で抽出できた課題を表し，補助グラフにてグループ化前後どちらがより良い評価を得られたかを表す．抽出可能な検索課題の割合は，グループ化前については 32%，グループ化後については 31%であることがわかる．これは，それぞれが異なる適合課題を同程度抽出可能であること，および抽出可能な検索課題が非常に少ないことを意味している．

この結果より，二つの動的スコアを併合することにより適合文書を多く抽出することが可能であると考えられる．

提案手法の各スコアがそれぞれ特徴あるスコアであるため，各スコアをすべて併合することにより，最終的なスコアを算出することを考える．最終スコア *Score(p)*の算出式を以下に示す．なお，*DynamicN(p)*，*DynamicG(p)*はそれぞれ文書*(p)*におけるグループ化前後動的スコアを，*Wdn*，*Wdg* は各スコアの重み係数を表す．

$$Score(p) = Wr \cdot Retrieval(p) + Wsn \cdot StaticN(p) + Wsg \cdot StaticG(p) + Wdn \cdot DynamicN(p) + Wdg \cdot DynamicG(p)$$

図 4に最終スコアリング評価結果を示す．結果より，(*Wr*，*Wsn*，*Wsg*，*Wdn*，*Wdg*)=(2,1,2,0,0)の場合に最もよい性能であることが確認できた．また，動的スコアを併合した場合には性能が下がっていることがわかった．

6. 考察

以上の結果より，全文検索スコアとグループ化前後の静的スコアを併合した場合の WRR におけるランク 100 時点での評価は，全文検索スコアのみを利用した場合に比べ+200%程度，PageRank アルゴリズムのみを利用した場合に比べ+10%程度，評価が向上するという結果が得られた．しかし，全文検索スコア，グループ化前後静的スコア，およびグループ化前後動的スコアをすべて併合した場合には，全体的に評価が低下するという結果になった．この評価低下の原因としては動的スコアリングの精度不足が考えられる．全文検索スコアとグループ化前後それぞれの動的スコアを併合した場合の評価について調査した結果，動的スコアを併合した場合の評価は，全文検索スコアのみによる評価に比べ低下していることがわかった．特に全文検索スコアとグループ化後動的スコアを併合した場合の評価が悪く，すべてのスコアを併合した場合の評価に多くの影響を与えていると考えられる．またグループ化手法にて高精度のグループ化が実現できていないことも，グループ化後動的スコアの評価が悪い原因として考えられる．

7. おわりに

本論文では，類似情報を持つ Web ページ群をグループ化することによりリンク構造上の隣接関係を拡張し，リンク構造解析スコアリングを静的，動的に算出するランキング手法を提案した．また提案手法について実験を行った．その結果，提案手法による精度向上を確認するとともに，グループ化手法を再検討する必要があることを確認した．

今後は，グループ化手法の再検討と，さらなる精度向上のためスコア併合式における各スコア重み係数の検証，スコア併合式自体の検証を行っていく．

文　献

[1] -, “Google” < http://www.google.com/ >

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

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

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

[5] 中窪仁, 佐藤隆士, “Web 検索におけるリンク構造解析を利用したランキング法,” DBWS2004, 信学技報 Vol.104 No.177, no.DE2004-65, pp99-103, Jul. 2004.

[6] T. Sato, T. Satomoto, and K. Han, “NTCIR-3 PAT Experiments at Osaka Kyoiku University,” In *Working Notes of the 3rd NTCIR Workshop Meeting Part III: Patent Retrieval Task*, pp.21-24, Tokyo, Japan, 2002.

[7] K. Eguchi, K. Oyama, A. Aizawa, and H. Ishikawa, “Overview of WEB Task at the Fourth NTCIR Workshop,” In *Working Notes of the Fourth NTCIR Workshop Meeting*, pp.ov1-ov2, Tokyo, Japan, 2004.

[8] K. Eguchi, K. Oyama, A. Aizawa, and H. Ishikawa, “Overview of the Information Retrieval Task at NTCIR-4 WEB,” In *Working Notes of the Fourth NTCIR Workshop Meeting*, pp.ov3-ov15, Tokyo, Japan, 2004.

[9] K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama, “Overview of the Web Retrieval Task at the Third NTCIR Workshop,” *NII Technical Report*, NII-2003-002E, 2003.

[10] 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*, vol.E86-D, No.9, pp.1804-1813, 2003.

表 1 グループあたり Web ページ数

最小値	1
最大値	30,446
平均値	5
中央値	1

表 2 グループ化前後比較（静的）

	グループ化前	グループ化後
ノード総数	23,670,000	4,500,000
リンク総数	79,700,000	18,140,000
スコア最大値	2.6126E-04	4.1985E-07
スコア最小値	7.3143E-09	3.3442E-08
スコア平均値	4.2231E-08	2.2230E-07
スコア中央値	8.3860E-09	2.2612E-07

表 3 グループ化前後比較（動的）

	グループ化前	グループ化後
ノード総数	192,500	124,041
リンク総数	95,848	120,292
スコア最大値	4.8634E-01	5.6874E-02
スコア最小値	6.8460E-05	7.6747E-05
スコア平均値	4.0000E-04	6.3694E-04
スコア中央値	7.0123E-05	5.1010E-04

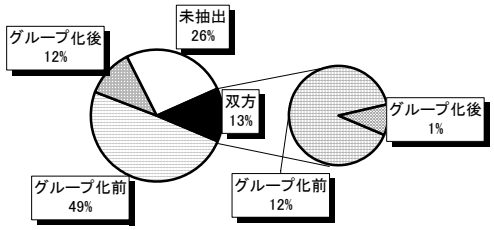

図 1 適合課題抽出割合（静的）

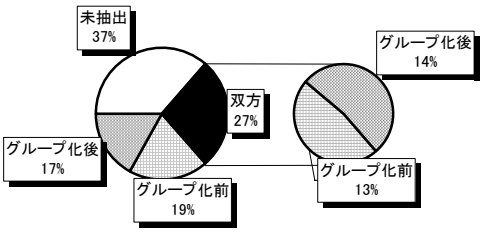

図 3 適合課題抽出割合（動的）

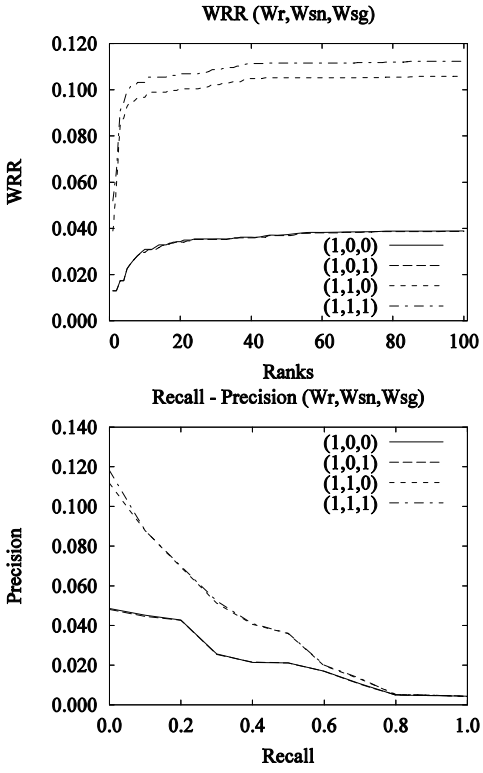

図 2 評価結果（静的スコアリング）

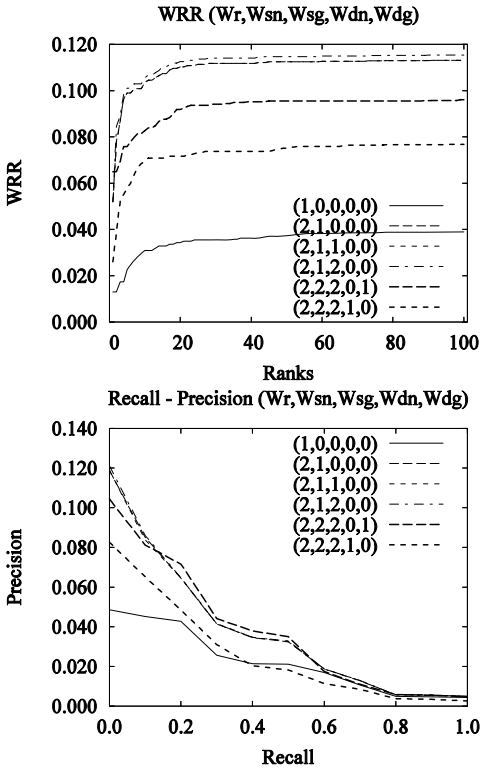

図 4 評価結果（最終スコアリング）