2004－DBS－134　(II)　(56)
2004／7／14

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

中窪 仁[†]　　佐藤 隆士[‡]

あらまし

インターネットが普及した現在，WWW 空間上には膨大な数の文書が存在している．本論文では，この情報群より必要な情報を抽出し，検索結果の上位にランキングする手法を提案する．はじめに関連研究を紹介する．次に我々が本論文で提案する手法として，Web ページのグループ化手法，Web ページのグループ化を用いた静的スコアリング手法，全文検索結果にリンク構造解析を適用する動的スコアリング手法について述べる．最後に，実験について述べる．

Ranking Method Using Link Structure Analysis in Web Retrieval

Hitoshi NAKAKUBO[†]　and　Takashi SATO[‡]

Abstract

A huge number of documents exist on the WWW space now. We propose a method of ranking a result of retrieval from them in important order. In the beginning, we introduce relation researches. Next, we describe Web Page Grouping Method, Static Scoring Method with Web Page Grouping, and Dynamic Scoring Method of Applying Link Structure Analysis to Full-text Search Result as a proposal method in this paper. Finally, we describe our experiment.

1. はじめに

インターネットが普及した現在，WWW 空間上には膨大な数の文書が存在しており，この中より必要な情報を抽出する事は非常に困難である．WWW 空間上より必要な情報を抽出するためのツールとして Web 検索システムが存在するが，検索精度が未だ十分でなく，必要とする情報に到達できないことも多々ある．一般的な Web 検索システムでは，ユーザより検索語句を受け取り，その検索語句を含む Web ページを検索結果として出力する手法を取り入れている．しかし Web ページ本文と検索語句に頼った情報抽出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いる．また，特定 Web ページに関連する Web ページ集合を抽出する手法も存在するが，この手法はリンク構造解析を利用しているものの精度的には十分とはいえない．

そこで本論文では，Web ページ本文と検索語

†大阪教育大学大学院総合基礎科学専攻
Course of Information Science, Osaka Kyoiku University
‡大阪教育大学情報処理センター
Information Processing Center, Osaka Kyoiku University

句を利用した検索結果に，リンク構造解析によるランキング結果を併合することによって，Web 検索精度を向上させる手法を提案する．リンク構造解析には Google[1]にて利用されている PageRank アルゴリズム[2]を用いることとし，以下の二つの手法を提案する．

1. グループ化した Web ページ集合にリンク構造解析を適用する，静的スコアリング手法
2. 全文検索結果集合にリンク構造解析を適用する，動的スコアリング手法

以下，2節にて関連研究を紹介する．3節にて提案手法として Web ページのグループ化手法，静的スコアリング手法，動的スコアリング手法について述べる．4 節にて実験方法について述べ，5 節にてまとめる．

2. 関連研究

2.1. PageRank アルゴリズム

PageRank アルゴリズムは，Web ページ間のリンク構造にランダムウォークモデルを適用し，WWW 空間上に存在する全 Web ページへの遷移確率を基にスコアリングを行う手法である．このスコアは WWW 空間上に存在する各 Web ページの特性を示す固定値となり，各 Web ページの被参照度を明確に示す値となる．

しかし，PageRank アルゴリズムは遷移確率を基にスコアリングしているため，「リンク構造上隣接関係にないが関連している Web ページ間」では Web ページ間で相互に与え合う影響が減少してしまう問題があると考えられる．

2.2. HITS アルゴリズム

HITS アルゴリズム[4]は，Web ページ間のリンク構造を解析し，検索語句に対して適切なコミュニティを抽出する．HITS アルゴリズムにおける“ authority ”とは，特定の検索語句に関する的確な情報を持つ Web ページ集合である．“ authority ”ページはリンク構造上隣接関係を持たず，“ hub ”ページを介して関係を持つ．これを「良質の authority は複数の良質の hub によってリンクされ，良質の hub は複数の良質の authority にリンクしている」と定義し，スコアリングおよび分類を行う．このスコアは Web 検索結果集合によって変動する可変値であり，Web 検索結果集合中での各 Web ページの有用度を示す値となる．

しかし HITS アルゴリズムには，抽出した Web ページに特定検索語句に関連のない Web ページへのリンクが大量に存在した場合などでは，ユーザが意図した Web コミュニティを抽出できるとは限らないという既知の問題がある．

3. 提案手法

3.1. 提案手法概要

提案手法を用いたランキング手順は，以下のようになる．

1. 全文検索用に検索対象全文書をインデクス化
2. 検索対象全文書中の全リンク構造を抽出
3. 2のリンク構造を元に Web ページをグループ化
4. PageRank アルゴリズムを3に適用
5. 検索語句にて全文検索およびスコアリング
6. 5の結果集合内のリンク構造を抽出
7. PageRank アルゴリズムを6に適用
8. 6のリンク構造を元に Web ページをグループ化
9. PageRank アルゴリズムを8に適用
10. 4，5，7，9の各スコアを併合

本論文では手順3, 8を Web ページのグループ化，手順4を静的スコアリング，手順7, 9を動的スコアリング，手順10をランキングと定義し，以下でそれぞれについて説明する．

3.2. Web ページのグループ化手法

我々は類似分野の情報を持つであろう Web ページ群をグループとして扱うことを提案する．我々が定義するグループとは，「同一の作成者が作成し，類似分野の情報を持つと思われる Web ページ群」である．「類似分野の情報は同一の親を持つ部分木に含まれている」と仮定した上で，この定義を満たすグループの作成手順を二通り示す．それぞれ，ディレクトリ構造に基づくグループ化手法とリンク構造に基づくグループ化手法である．例を図 1に示す．例内の Web ページ名 A…E は同一の Web ページを示す．

A) ディレクトリ構造方式

1. URL より各 Web ページのディレクトリ構造を抽出する．
2. ディレクトリ内の各 Web ファイルを，ディレクトリ内の index.html ファイ

ルとして統合する．

3. 各 Web ページ中のリンク構造を，グループへのリンクに置換する．

B) リンク構造方式

1. 全 Web ページのリンク情報より，Web ページを Web サイト単位に分割する．Web サイトとは同一人物によって作成された一連の Web ページ群であると定義する．
2. 各 Web サイトの入り口となるトップページを決定し，トップページを根とする木構造を構築する．
3. 葉にあたる Web ページを親の節に統合する．
4. 各 Web ページ中のリンク構造を，グループへのリンクに置換する．

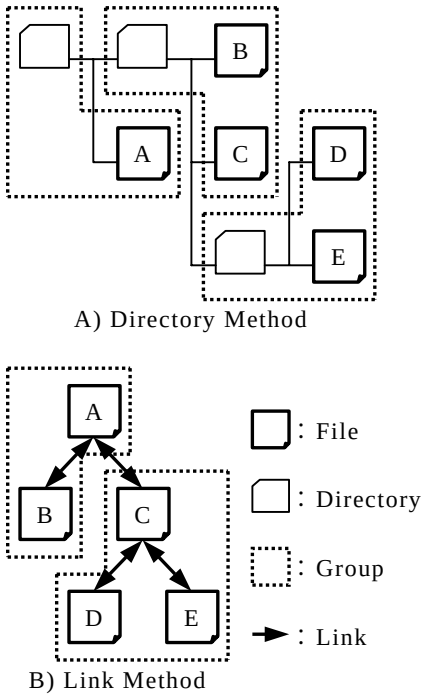

図 1 Web ページグループ化例
Fig. 1 Example of Web Page Grouping

ディレクトリ構造方式には，リンク構造解析が不要であるためグループ化処理が簡単であるという長所がある．反面，作成者がディレクトリ分類を正確に行っていない場合には類似分野の情報を持つグループを抽出できない可能性があるという短所がある．リンク構造方式には，作成者が意図した分類どおりにグループ化されるため，正確なグループが抽出可能であるという長所がある．反面，リンク構造解析を利用したグループ化は非常に難しいという短所がある．リンク構造解析を利用した Web ページグループの抽出手法については多くのものが提案されているが，未だに素晴らしい手法が提案されていないのが現状である．

3.3. 静的スコアリング手法

我々は，PageRank アルゴリズムを全リンク構造に対して適用するのではなく，グループ化済みのリンク構造に適用する手法を提案する．

以下図 2にて，静的スコアリングの例を示す．例中の Web ページ A…E は図 1と同一のものであり，Web ページ F…H は Web サイト外部の Web ページである．

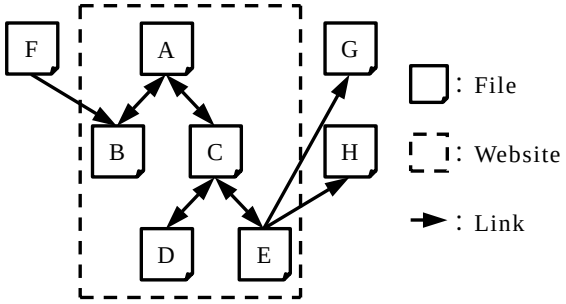

図 2 静的スコアリング例
Fig. 2 Example of Static Scoring

図 2の例について，グループ化なし，ディレクトリ構造方式グループ化，リンク構造方式グループ化を行い，PageRank アルゴリズムを適用した結果を表 1に示す．

表 1 静的スコア例
Table 1 Example of Static Score

Grouping Method	Page							
	A	B	C	D	E	F	G	H
No	0.26	0.15	0.26	0.11	0.11	0.00	0.06	0.06
Dir.	0.22	0.38		0.22		0.00	0.09	0.09
Link	0.22		0.36			0.00	0.22	0.22

この結果より，グループからリンクされている Web サイト外部の Web ページの PageRank スコアが増加していることがわかる．これは，Web ページをグループ化することにより，グループ

内のリンク構造上隣接関係が無視できるようになり，2.1節で挙げた PageRank アルゴリズムの問題が一部解消できたためと考えられる．またこの例だけでは検証できないが，グループ化の範囲を「同一人物により作成された一連の Web ページ」に制限することで，2.2節で挙げた HITS アルゴリズムの問題も解消できると考えられる．

グループ化の弊害として，本来のリンク構造が置換され各 Web ページの特性が失われるため，グループされた Web ページの PageRank スコアが均一化されてしまっていることが挙げられる．

3.4. 動的スコアリング手法

次に我々は，全文検索結果集合に対して PageRank アルゴリズムを適用する手法を提案する．全文検索結果集合は，特定検索語句を含むコミュニティである．我々は，このコミュニティ内でリンク構造解析を行う事により，コミュニティ内での重要な情報を特定できるのではないかと考える．そこで，「検索結果集合にリンク元 Web ページとリンク先 Web ページが両方含まれるリンク構造」のみを抽出し，PageRank アルゴリズムを適用する．また，検索結果集合は検索対象集合に比べ非常に小さな集合となるため，集合内にリンク構造が含まれない可能性もある．この問題を解消するため，「検索結果集合にリンク元 Web ページグループとリンク先 Web ページグループが両方含まれるリンク構造」に PageRank アルゴリズム適用範囲を拡張する手法をとる．

以下図 3にて，動的スコアリングの例を示す．

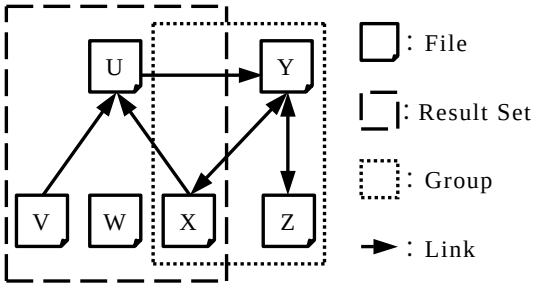

図 3 動的スコアリング例
Fig. 3 Example of Dynamic Scoring

図 3の例について，全文検索結果集合に PageRank アルゴリズムを適用した場合，全文検索結果集合をグループまで拡張して PageRank アルゴリズムを適用した場合の結果を表 2に示す．

表 2 動的スコア例
Table 2 Example of Dynamic Score

Grouping	Page					
	U	V	W	X	Y	Z
No	1.00	0.00	0.00	0.00	0.00	0.00
Yes	0.50	0.00	0.00	0.50		

全文検索結果集合をそのまま利用した場合，リンクされている Web ページが Web ページ U のみであるため，PageRank スコアも Web ページ U にのみ与えられている．しかし，適用範囲を拡張した場合，リンク構造の置換により Web ページ X へもリンクがされていることになり，PageRank スコアが与えられていることがわかる．これらの PageRank スコアは検索語句によって変動する値であり，検索結果集合に含まれる各 Web ページの特性を明確に表す値になると考えられる．二つの PageRank スコアはそれぞれ別のねらいによって算出された値であり，これらを3.5節に示す方法で併合することにより動的スコアと扱う．

この手法の問題点として，検索語句それぞれについてリンク構造解析を行い，PageRank アルゴリズムを 2 回適用するため，処理時間がかかってしまうことが挙げられる．

3.5. ランキング

静的スコアリング，動的スコアリングおよび全文検索スコアを加味することで，最終的なランキングを行う．

全文検索スコアは検索語句に特化したスコアリングがなされており，これをランキングの主に据えることとする．静的スコアは全 Web ページの特性を大まかに捉えたスコアであり，動的スコア二つはそれぞれ検索結果集合内での被参照度を明確に示すスコア，検索結果集合を拡張認識した場合の被参照度を示すスコアである．これらのスコアの特性を加味した上で，最終的なスコア算出式を以下のように定義する．

$$Score(p) = w_r\, Retrieval(p) + w_s\, Static(p) + w_d\, Dynamic(p)$$

w: Weight $(w_r > w_d > w_s, w_{d1} > w_{d2})$

$Retrieval(p)$: Full - text Search Score of Document p.

$Static(p)$: Static Score of Document p.

$D_1(p)$: Dynamic Score of Document p without Grouping.

$D_2(p)$: Dynamic Score of Document p with Grouping.

$Dynamic(p) = w_{d1} D_1(p) + w_{d2} D_2(p)$

各スコアに付ける重みの値 w については，今後検討していく必要がある．

3.6. 考察

我々の提案したランキング手法は，静的スコアリングおよび動的スコアリングにてコミュニティの抽出を実現可能である．また，動的スコアリングおよび全文検索スコアにて検索語句を含む情報を上位に押し上げるスコアリングが可能である．この二つの要因より，ランキング時の重みの最適値が決定すれば非常に興味深い結果が得られ,検索精度が向上すると考えられる．我々の手法ではグループ化範囲を制限することにより2節で挙げたような問題の発生を最低限に抑えることも可能であると思われる．

提案手法の問題点としては，スコアリング処理が従来のものよりも多いために処理時間がかかるということが考えられる．

4. 実験概要

我々が今後行っていく実験の概要を以下に述べる．

実験用検索対象には, NTCIR[5][6]にて用意されている 100Gbyte のデータ（NW100G-01）を用いる．これには約 1100 万件の Web ページが格納されており,リンク総数は約 8000 万件である.検索語句についても NTCIR にて用意されたものを使用する．

実験環境としては,CPU が Pentium4 2.4GHz, メモリが 1GByte のハードウェアを使用し，FreeBSD を OS として動作させる．

全文検索システムには本研究室で作成したグラムベースインデクスを用いた検索システム[7]を利用する．全文検索結果のスコアリング法には *tf・idf* 法を使用する．グループ化には簡単のため，ディレクトリ構造に基づく手法を用いる．静的スコアリングに関してはグループ化前後の PageRank スコアを算出する．動的スコアリングについては,全文検索結果集合 5000 件を対象に PageRank スコアを算出する．これらのスコアについて, 3.5節のスコア算出式を用いて最終的なランキングを決定する．その際，重みに関する検証も同時に行う．

評価には，NTCIR で使用されている評価方式を使用する．

5. おわりに

本論文では，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として，Web ページのグループ化手法，静的スコアリング手法，動的スコアリング手法について提案した．今後は提案内容について実験を重ね，精度向上に関する調査, ランキング時の重み最適値の検討, Web ページのグループ化手法の検討を行っていく．

文　献

[1] http://www.google.co.jp/.

[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/pagerank sub.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] 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.

[6] 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.

[7] 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.