Webページの重要度ファクタに関する一考察

○ 福島 俊一
松田 勝志
高野 元

A Study of Page Ranking Factors for WWW Search Engines

○ Toshikazu Fukushima
Katsushi Matsuda
Hajime Takano

This paper surveys page ranking factors used in the current WWW search engines, such as (1) relevance to query keywords, (2) freshness, (3) popularity, (4) citation rank and (5) page types. The relevance to query keywords have been studied in the traditional information retrieval researches. However, other factors are introduced into the WWW search engines in order to improve their ranking performance, because WWW contents are heterogeneous and changeable large-scale hypermedia. The freshness, the popularity and the citation rank are the factors introduced from a viewpoint of contents reliability. On the other hand, the relevance to query keywords and the page types are the ones corresponding to user's domain and task in problem solving. Selection and combination of these factors must be refined for satisfying user's information needs.

1. はじめに

インターネット上のWWW (World-Wide Web)から必要な情報を見つけ出すために、WWWサーチエンジン[1]の利用が広まっている。WWW サーチエンジンはまさに情報検索の研究成果を生かすべき格好のターゲットになり得るものであるが、伝統的な情報検索研究で扱われてきた文献資料(論文/新聞/特許など)と比較して、WWW はかなり異なる様相を呈している。すなわち、WWWコンテンツは次のような特徴を持っている。

- **大規模ハイパーメディア** ： ページと呼ばれるドキュメント単位がリンクで結ばれ、クモの巣のような構造を成している。ページは、テキストのみでなく、画像・音声データなども含むマルチメディアドキュメントである。ページの数は既に、国内で数千万ページ、全世界では数億ページに達していると言われる。
- **ヘテロ** ： WWW コンテンツの種類(タイプ)は様々で、カタログ・新聞記事・リンク集など異質なページが混在している。情報発信者も多様で、公式ページから個人メモ風のものまで不均質である。使われている言語も多種類である(日本語・英語ほか)。
- **日々変化** ： 日々、新しいページが増えている。古いページの更新も頻繁に発生し、リンク先のページがいつの間にか変わっていたり、時にはなくなっていることもある。

このような特徴を持つ WWW を対象とするため、WWW サーチエンジンでは、検索結果のランキング(重要度計算)において、伝統的なランキング手法にはなかった新しいファクタが導入されている。そこで、本稿では、現在の WWW サーチエンジンにおいて Web ページの重要度計算にどのようなファクタが使われているかを概観する。そして、それらのファクタの持つ視点やファクタ間の関係について考察する。

2. Webページの重要度ファクタ

現在の WWW サーチエンジンにおいて Web ページの重要度計算に用いられているファクタとし

て、以下のような 5 つが知られている。伝統的な情報検索研究では、検索語との適合度がランキングにおける中心的なファクタであった。それに対して、Web ページのランキングでは、ページ更新日時(新鮮度)、参照履歴(人気度)、リンク構造(引用度)、ページタイプなどの新しいファクタが導入され、有効に作用している。第1節で述べたWWWコンテンツの特徴と対応付けると、引用度は「大規模ハイパーメディア」という特徴、人気度・引用度・ページタイプは「ヘテロ」という特徴、新鮮度は「日々変化」という特徴を各々反映していると考えられる。

(1) 検索語との適合度

「そのページの主題に検索語が適合しているページほど重要」という考えに基づく。情報検索におけるランキングの最も基本的な考え方で[2]、多数の WWW サーチエンジンで採用されている。

この考え方による代表的な重要度計算法として TF*IDF スコアがある。TF 値(term frequency)は、ページ内に出現する検索語の数をページサイズ(テキスト長)で正規化したもので、検索語が多数出現するページほど重要だという考えを表す。一方、IDF 値(inverted document frequency)は、検索語の出現するページ数の総ページ数に対する割合の逆数で、ありふれていない検索語の出現するページの方が重要だという考えを表す。TD*IDF スコアは、これら TF 値と IDF 値をかけ合わせたもので、ありふれていない検索語が多数出現するページほど重要ということになる。

また、Web ページのような HTML ドキュメントを対象とした場合には、HTML タグを考慮することで、タイトルや見出し部分に検索語が出現した場合にスコアに加点するような手法も有効である。

(2) ページ更新日時(新鮮度)

「新しいページほど重要」という考えに基づく。同じような内容を取り上げていても、新しいページの方が、新たに得られた情報や議論の結果を取り込んで情報の価値・量が高まっている可能性が高い。また、例えばイベント案内や求人情報のように、利用者が WWW から探そうとする情報自体が、鮮度に価値を有する類のものが多いという側面もある。

新鮮度のスコアは、Web ページの更新日時が新しいほど高くなる。情報鮮度を重視した WWW サーチエンジンとして FreshEye (http://www.fresheye.com/)[3] があるほか、NETPLAZA (http://netplaza.biglobe.ne.jp/keyword2.html)などでもページ更新日時の新しい順に検索結果のソートができる。

(3) 参照履歴(人気度)

「多数の利用者が参照したページほど重要」という考えに基づく。

米国のWWWサーチエンジンHotBot (http://www.hotbot.com/)の採用したDirectHit Popularity Engine (http://www.directhit.com/)がよく知られている[4]。

人気度は過去の参照履歴から算出できる。すなわち、ある検索語とその検索結果から利用者が選択したジャンプ先ページ(URL)の組を記録しておけば、各検索語に対してどのページ(URL)へのジャンプ回数が多かったかがわかる。この過去のジャンプ回数が、検索語ごとの各ページの人気度に相当する。

(4) リンク構造(引用度)

「多数引用されるページは重要」および「重要なページに引用されるページも重要」という考えに基づく。このような考え方は、従来も論文の質の評価などに用いられていた。WWW のハイパーメディア構造とリンクの使われ方に着目すれば、WWW コンテンツに同様の考え方を適用しようというのは自然な発想である。

米国スタンフォード大学で開発された WWW サーチエンジン Google (http://www.google.com/ あるいは http://google.stanford.edu/)では、本稿でいう引用度に相当するページランクというスコアを次式で計算している[4][5]。

$$PR(A) = (1-d) + d(PR(T1)/C(T1) + \cdots + PR(Tn)/C(Tn)) \qquad d = 0.85(\text{定数})$$

PR(X)： ページ X のページランク値

T1, … , Tn： ページ A を引用している(A へリンクを張っている)ページが n 個

C(X)： ページ X が引用しているページ数(X の持つ外部へのリンク数)

この式によれば、各ページのページランク値をそのページから出ていくリンク数で割ったスコア値をリンク先ページへ伝播させ、リンク先では入ってくるスコア値の合計をそのページのページランク値とする、という計算を再帰的に実行することで、上述の考えを実現している。

また、NETPLAZA (URL は前出)においても、引用されるページ数(被リンク数)を重要度とみなしたランキング機能を提供している。

(5) ページタイプ

「利用者の問題解決に合ったタイプのページほど重要」という考えに基づく。例えば、パソコン購入という問題解決タスクには「カタログ」タイプ、就職や転職という問題解決タスクには「求人情報」タイプというように、問題解決タスクの各々に適した特定のページタイプが存在する。

筆者らは、このようなページタイプは各々、固有のスタイルを有すると考え、Web ページ(HTML ドキュメント)の構造的な特徴に着目したページタイプの自動分類法を開発した[6]。この分類手法では、各ページタイプの典型を意味する特徴記述を用意しておき、その特徴をどれくらい満足しているかを、各ページタイプに関するスコアとして算出する。

筆者らは現在までに、「カタログ」「リンク集」「求人情報」「プレゼント情報」「掲示板/チャット」という 5 種類のページタイプを扱う分類システムを開発し、NETPLAZA (URL は前出)において公開している。NETPLAZA は、ページタイプを絞り込み条件に用いることができる初めての WWW サーチエンジンである。

3. 重要度ファクタ個別の問題点

第 2 節で述べた重要度ファクタの各々は、以下のような問題を有している。

- **検索語との適合度の問題** : 利用者にとって情報要求を検索語だけで的確に表現することは極めて難しい。特に WWW サーチエンジンの利用者が入力する検索語の数は、多くの場合 1～2 語に過ぎない。ヘテロな WWW コンテンツを対象として、単一の適合度スコア計算法が通用するかは疑問である。
- **新鮮度の問題** : 同じような内容のページであれば、新しい方が高い価値を持つという傾向は確かにあるが、常に新しいものが望まれるとは限らない。例えば、チャットのようなページが上位にくることがある。
- **人気度の問題** : 多数の利用者に人気があるページは、大まかには個人にとっても価値が高いという仮定は妥当であるが、誰にでも成立するものとは限らない。その WWW サーチエンジンの利用者の多数決的な価値観を反映したものになる。また、十分な統計量の得られない検索語に対しては使えない。
- **引用度の問題** : 検索語が出現してさえいれば、どのような文脈で出現しているかに無関係に、引用度スコアの高いページが上位にくる。
- **ページタイプの問題** : 検索語が出現してさえいれば、どのような文脈で出現しているかに無関係に、典型的なページタイプのスタイルを持つページが上位にくる。例えば、FAX モデムのカタログが欲しいとき、「FAX モデム」という検索語が出現したページ群を「カタログ」タイプらしさでランキングすると、FAX モデムを内蔵したパソコンの典型的な「カタログ」ページが上位にきてしまう可能性は十分にある。

4. 重要度ファクタ間の関係

WWW サーチエンジンを利用者の問題解決のための手段ととらえてみると、第 2 節であげた 5 つのファクタは、次のような関係にあるとみなせる。

まず、利用者の問題解決についてドメインとタスクを分けて考えると、検索語は主に問題解決のドメインを絞り込むためのファクタとなり、ページタイプは主に問題解決のタスクを絞り込むためのファ

クタとなる。一方、新鮮度・人気度・引用度は、利用者の個々の問題解決への適合性とは独立した、WWWコンテンツの信頼性を評価するためのファクタである。

問題解決のドメインとタスク、および、WWWコンテンツの信頼性は、基本的に独立した3つの評価軸であろう。したがって、利用者の問題解決のための情報要求に適合するようなWebページのランキングを実現するためには、これらを適切に組み合わせた総合的な重要度(＝問題解決ドメイン適合度×問題解決タスク適合度×WWWコンテンツ信頼性)の計算が必要になるはずである。第3節において個別ファクタでのランキングの問題点を指摘したが、これは1つの評価軸のみによるランキングの限界を示しているものと考えれば納得ができる。

しかし、3つの評価軸の適切な組み合わせ方や5つの重要度ファクタの適切な選別は、必ずしも簡単な問題ではない。各評価軸や各ファクタにどのような重みを与えるかの指針は明らかになっていないし、複数の評価軸あるいはファクタを組み合わせた結果として、利用者には理解しにくいような重要度スコアになってしまう恐れもある。

さらに、本節で整理した評価軸をベースに考えてみると、WWWコンテンツの信頼性に関する軸については、情報発信者(WWWページの作成者)が誰かというファクタも考慮すべきかもしれない。また、利用者側の軸については、問題解決のドメインとタスクのほかに、利用者プロファイルも関わってくるものと思う。

5. おわりに

現在のWWWサーチエンジンにおいてWebページの重要度計算に用いられている5つのファクタ——検索語との適合度、新鮮度、人気度、引用度、ページタイプ——を概観し、それらが利用者の問題解決のドメインとタスク、および、WWWコンテンツの信頼性という視点に対応していることを述べた。利用者の問題解決のための情報要求に適合するようなWebページのランキング(重要度計算)を実現するためには、今後、これらのファクタの適切な選別と組み合わせに関して検討していくことが必要である。

参考文献

[1] 原田、「サーチエンジン徹底活用術」、オーム社、1997年。

[2] D.Harman, "Ranking Algorithms", Information Retrieval: Data Structure and Algorithms (edited by W.B.Frakes, R.Baeza-Yates, Prentice Hall, 1992), pp.363-407.

[3] 住田・鈴岡・平野・野上、「WWW 情報フィルタリング・検索システム(FreshEye)－サービス概要－」、情報処理学会第57回全国大会3L-5、1998年10月。

[4] 橋本、「サーチエンジンの新しいテクノロジーとアクセス向上」、アクセス向上委員会通信(http://www.access.or.jp/)、Vol.62、1998年11月16日。

[5] S.Brin and L.Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", http://google.stanford.edu/long321.htm

[6] 松田・福島、「文書タイプ分類による問題解決向き WWW 検索システムの開発と評価」、情報処理学会情報学基礎研究会53-2、1999年3月1日。

福島 俊一　NEC ヒューマンメディア研究所 (〒630-0101 奈良県生駒市高山町8916-47)
松田 勝志　同上
高野 元　NEC C&C メディア研究所 (〒216-8555 神奈川県川崎市宮前区宮崎4-1-1)
Toshikazu Fukushima (fuku@HML.CL.nec.co.jp) Human Media Research Labs., NEC Corp.
Katsushi Matsuda (mat@HML.CL.nec.co.jp) 同上
Hajime Takano (gen@ccm.CL.nec.co.jp) C&C Media Research Labs., NEC Corp.