情報構造解析を用いたWWW検索ランキング方式

3P－7

高野 元　久保 信也
NEC ヒューマンメディア研究所

1.はじめに

本稿では、WWW 検索を対象とした、情報構造解析を用いたランキング方式について述べる。従来の WWW 検索は膨大なページを均一なドキュメント集合と考えるのに対して、本稿で提案する方式は、ホスト・ディレクトリ名だけでなくリンク構造を用いて判別した情報構造に基づいたランキングシステムを提供する。これにより、重要なサイトおよびサイト内の手がかりになるページを積極的にユーザに提示することで膨大な検索結果のブラウジングが効率化できる。また、本方式を実際の WWW 検索システムに適用した結果を紹介する。

2.WWW 検索の改善方針

WWW 検索の問題点は、単に検索結果が膨大になるということではなく、「見るべき情報を、少ない手間で選別できる」ようになっていないことであると考える。すなわち、

- 重要なページが必ずしも検索結果の上位に出てこない。また、重要さの定義があいまいなために、ユーザの混乱を招いている。
- 検索結果が WWW サイトの構造に無関係に出力されるため、何度も同じサイト上のページをチェックしなければならない。
- 検索結果に見るべきページが含まれていなかったり、不要なページが大量に含まれる。

といった問題点の解決が必要である。これに対して、以下の処理を導入して解決を図る。

- 検索語への適合度ではなく、認知度の強いものが重要であるとし、リンク参照を用いたページランク解析を導入する。
- リンク参照や URL 構造を用いて情報構造を判定するインフォメーション・ユニット解析を導入する。ここで、たとえばユーザ・ホームページや企業の製品ホームページなど、ユニット=ホストではない点に注意。
- 上記ページランクとインフォメーション・ユニットに基づいて、検索結果をサイトでクラスタリング・ソートし、重要ページの追加・不要ページの削除をする、リンク構造要約を導入する。

3.情報構造解析を用いたランキング方式

前節で挙げた処理アルゴリズムの概要を説明する。

"A Ranking Method for WWW search using Information Structure Analysis," Hajime Takano and Nobuya Kubo, Human Media Research Labs., NEC Corporation
E-Mail: {gen, nobuya}@hml.cl.nec.co.jp

3.1.ページランク解析

ページランクは、そのページが WEB 空間においてどの程度の重要度を持つかを示す指標である[1]。ここでは、さらに電子ニュースなどでの参照数や利用履歴も考慮できるようモデルを拡張している(図1参照)。詳細は省くが、ページ u におけるページランク $R(u)$は、次式のように定義できる。

$$R(u) = \sum_{v \in B_u} \frac{W_{v \to u}}{W_v} \bullet R(v) + E(u) \quad (1)$$

ここで、Buはページuへのリンクを持つページの集合、$W_{v \to u}$はページvからuへのリンク重み、Wvはページvから出て行くリンクの重みの総和数である。つまり、リンク重みとは、ページの重要度をどのようにリンク先ページに分配するかを表している。、$E(u)$はページuをナビゲーションの起点として選択する確率である。よって、ページに対して式(1)を満たすRを算出できれば良い。

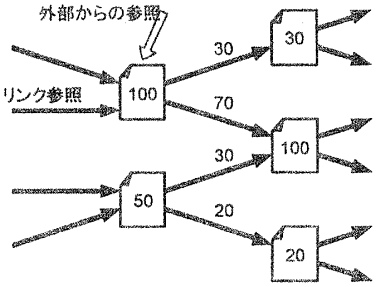

図 1 ページランクのモデル

3.2.インフォメーション・ユニット解析

インフォメーション・ユニットとは、(1)代表ページを持ち、(2)ユニットの範囲があり、(3)主要なページ関係は木構造をなす、という特徴を持つページ集合である(図2参照)。

こうしたユニットを定義するために、

- ユニットの代表ページは、同一ホスト上で外部または内部からのリンク参照数が多いものを選出し、
- 代表ページのホスト名/ディレクトリ名を用いてユニットに属するページの範囲を設定し、
- 代表ページを起点とするユニット内ページの主要なリンク構造を木構造として発見

するアルゴリズムを適用する。

3.3.リンク構造要約

リンク構造要約は、与えられたページ集合に対する一種のソートである。通常のソートと異なり、インフォメーション・ユニットの構造にしたがって、

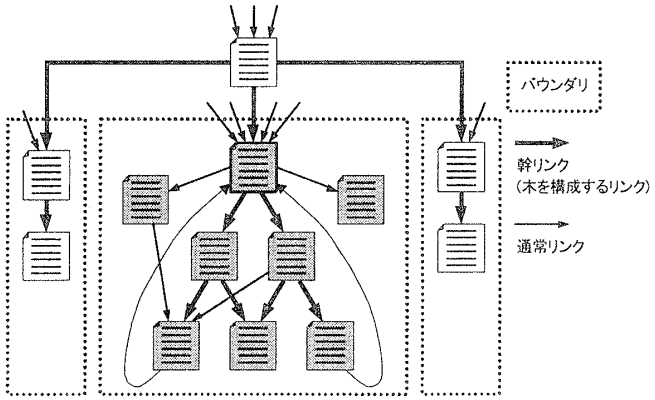

図 2 インフォメーション・ユニットのモデル

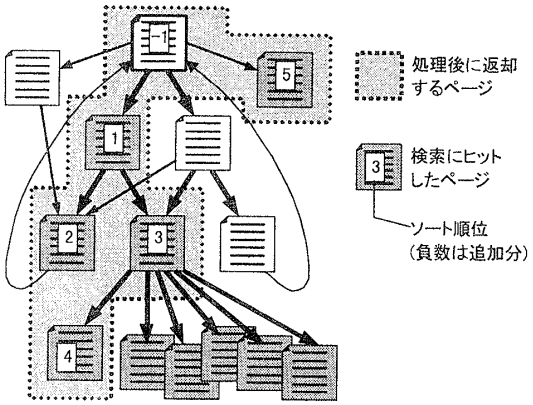

図 3 リンク構造要約のモデル

- ユニットごとにクラスタリング
- ページランクを用いて得るユニット重要度でソート
- ユニットごとに、検索結果に含まれていない重要ページ(代表ページなど)を挿入し、不要ページを削除

して、検索結果を整える(図 3 参照)。

4.システム評価

Web クローラ[2]が収集したページのリンク情報を格納したリンク DB を構築し、その上に前節で述べたアルゴリズムをインプリメントした。さらに収集したページは事前に全文検索エンジン[3]に登録しておき、この検索結果に LSS の機能を適用した結果を HTML 形式で出力する CGI プログラムを用意した。今回は、国内の Web 約 120 万ページに対して適用した。

図 4 に示すように、ユニットの代表ページのリストとして検索結果が出力される。ここで、ユニットをあらわすアイコンをクリックすると、ユニット内でキーワードにヒットしたページが現れるようにしている。この例では、キーワード「おかあさんといっしょ」の結果を示しているが、ホストの

ルートディレクトリではなく、ユーザディレクトリ以下のサブディレクトリがユニットの代表ページとして設定されていることがわかる。

5.おわりに

リンクなど WWW の構造情報を用いて、ページランクならびにインフォメーション・ユニットを求めることで検索結果のブラウジング効率を向上させる方式について述べた。今後は、各アルゴリズムの精緻化、ならびにコンテンツ解析との融合を図ることで精度の向上を図っていく。

参考文献

[1] "The Anatomy of a Large-Scale Hypertextual Web Search EngineSergey," Sergey Brin and Lawrence Page, pp107-117, Proceedings of 7th WWW Conf., May 1998

[2] "Development of a Scalable Web Crawler," Hajime Takano and Nobuya Kubo, pp334-339, NEC Res. & Develop.,Vol.40, No.3, July 1999

[3] "高速全文検索のためのフレキシブル文字列インバージョン法," 赤峯亨、福島俊一, pp35-42, ADBS'96, Dec. 1996

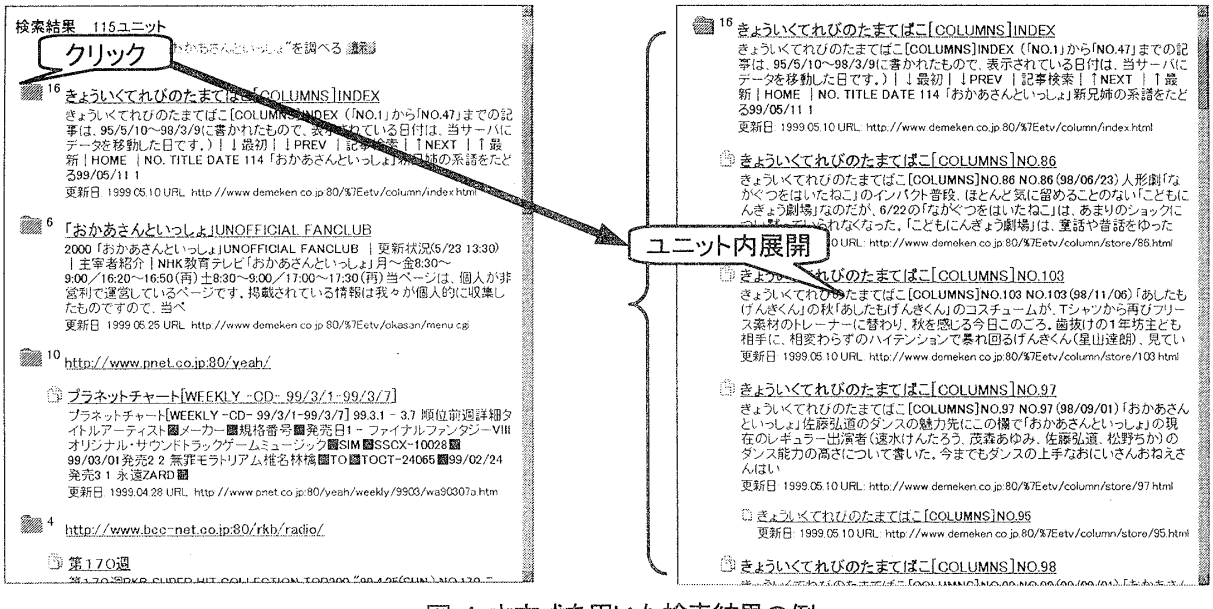

図 4 本方式を用いた検索結果の例