データベースシステム 120－2
（2000．1．24）

サイテーション・エンジン：
リンク解析を用いたWWW検索ランキングシステム

高野 元　　久保 信也
NEC ヒューマンメディア研究所

本報告は、WWW検索用のランキングシステムであるサイテーション・エンジンの機能と構成について述べる。現在のWWW検索の問題は、一つ一つのページを独立した文書として扱っていることであると考え、ユーザに提示すべき検索結果は個々のページの重要度ではなく、WWW サイトの構成を反映したものにすべきと考えた。サイテーション・エンジンは、WWW ページ間のリンク構造を解析することによって、上記機能を実現する。サイテーション・エンジンは WWW クローラが出力するリンク情報をデータベースに格納し、これを解析する。解析機能は、リンク参照関係を用いた重要度(ページランク)計算機能と、WWWページ間の関連を用いた情報構造(インフォメーション・ユニット)解析機能からなる。さらに、ここで得られたページランクとインフォメーション・ユニット情報に基づいて、検索結果を並べなおすリンク構造ソート機能を備える。これにより、全文検索エンジンと連携して、高度なランキング機能を備えたWWW検索システムを構成できることを、プロトタイプシステムによって確認した。

Citation Engine:
A Ranking System for the WWW Search Engine using Link Analysis

Hajime Takano　　Nobuya Kubo
Human Media Research Laboratories, NEC Corporation

This report describes functions and configuration of a ranking system for the WWW search engine, which we call "Citation Engine". We have figured out the problem in the WWW search system that the system treats each WWW page as an independent document. Therefore, the ranking method in the system should consider a structure of pages in a site. According to this consideration, "Citation Engine" was designed to give better ranking by using link analysis techniques. "Citation Engine" stores and analyzes whole link structures that are fetched by the Web Crawler system. There are two main analysis functions: (1) the page rank analysis, and (2) the information unit analysis. It also provides the structural sorting function, which organizes search result with information of the page rank and the information unit. The WWW search engine which gives well-organized search results can be build by integrating "Citation Engine" and the full-text search engine. Efficiency of "Citation Engine" is verified on a prototype system.

1. はじめに

WWW はインターネット上のあらゆるサービス基盤として利用されるに至り、世界中の個人や団体が情報発信を行うようになった。最近の報告では、全世界で 8 億ものページが存在するといわれており[Lawren99]、また日本国内に限っても 2 千万ページ以上存在するものと考えられ、検索サービスは WWW の利用上不可欠なものとなっている。

これらの検索エンジンサービスは、できるだけ多くの WWW ページを収集し、できるだけ多くの検索要求をこなすことが求められたため、スケーラビリティが開発における焦点となっていた。

しかし検索サービスの大規模化にともない、検索結果の数が数百件から数千件以上となることも多く、テキスト解析に基づく従来の検索技術とは異なる視点の導入が求められるようになった。

われわれは、現在のWWW検索における最大の問題は、一つ一つのページを独立した文書として扱っていることであると考え、ユーザに提示すべき検索結果は個々のページの重要度ではなく、WWW サイトの構成を反映したものにすべきと考えた。具体的には、ユーザが見るべきサイトの発見と、サイト内のナビゲーション・ブラウジングを助ける情報だけを提示する機能を提供する

こととした。

サイテーション・エンジンは、WWW ページ間のリンク構造を解析することによって、上記機能を実現する WWW 検索用のランキングシステムである。サイテーション・エンジンは Web クローラシステムと連携して、そこから出力されるリンク情報をデータベースに格納し、これを解析する。解析機能は、リンク参照関係を用いた重要度(ページランク)計算機能と、WWW ページ間の関連を用いた情報構造(インフォメーション・ユニット)解析機能からなる。さらに、ここで得られたページランクとインフォメーション・ユニット情報に基づいて、検索結果を並べなおすリンク構造ソート機能を備える。これにより、全文検索エンジンと連携して、高度なランキング機能を備えた WWW 検索システムを構成することができる。

本報告書は、サイテーション・エンジンの機能と構成について述べる。まず、2 章では開発の背景を述べ、3 章ではサイテーションエンジンの機能、特に解析アルゴリズムについて説明する。4 章でインプリメンテーションの内容を説明する。5 章では、初期評価、関連研究などの議論を行い、6 章でまとめる。

2. 開発の背景

2.1. WWW 検索システムの課題

WWW はホスト単位で情報を管理するため、広範囲のホスト上のページに対して検索をかけるためには、どうしてもそれらのページを一旦集めて集中管理する必要がある。したがって、WWW 検索サービスの構築を行うためには大きく分けて、(1) WWW 情報の収集、(2) 収集した情報のデータベース化、(3) データベースの検索(UI 含む)、の三つの機能が必要である。

われわれはこの(3)に対して、特にランキングシステムのあり方を議論し、次のような WWW 検索の問題点を解決すべきと考えた。すなわち、

a) 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で、ユーザの混乱を招いている。
b) 検索結果が WWW サイトの構造に無関係に出力されるため、ブラウジングの手間がかかる。
c) 見る必要のない関連ページが大量に検索結果に含まれるため、ブラウジングの手間がかかる。

2.2. ランキングシステムの要件

上記課題を解決する要件として、

- 適切なページ重要度の指標を導入
- WWW サイトの構造を反映した検索結果のクラスタリング
- 重要ページの追加および不要ページの削除

の方式について議論する。

2.2.1. ページ重要度

従来から使用されてきた検索語との適合度を用いたランキング方式は、特許や論文といったある程度均質な文書の検索には効果を発揮した。しかし WWW 上のページは、多数のトピックスを含むものから、複数で一つの形をなす断片的なものまで玉石混交であり、従来とは異なったページ重要度の概念が必要である。

適合度とは異なるページ重要度の指標には、

- 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(タスク合致度)

を挙げることができるが[Fuku99]、サイテーション・エンジンでは WWW の社会メディアとしての側面を重視し、情報発信力の強さをページ重要度の指標として採用することとした。

情報発信力とは、他からどれだけ認知されているかということであるため、重要なページから参照されているページは重要であるとする引用度を用いた重要度指標を導入する。

2.2.2. クラスタリング

従来の WWW 検索ではページを検索対象とするために、同一サイト上の異なるページが多数検索結果に含まれることがよくあった。しかし、同一サイト上のページは同一のテーマで編集されているため、ある 1-2 ページがユーザの検索目的に合致しなければ、そのサイト上の他のページも合致しないと考えるのが自然である。

したがって、同一サイトのものはクラスタリングすることで検索結果のブラウジング効率は向上すると考える。このとき気をつけなければならないのは、サイト=ホストとはならないことである。たとえば、プロバイダのユーザ・ホームページや企業の製品ホームページなどは、ルートディレクトリではなくある程度の深さのディレクトリに存在することがあるので、こうした構造も考慮したクラスタリングが必要である。

2.2.3. ページの追加・削除

検索結果のブラウジングに手間がかかるのは、最終的にそのページを開いて周辺をナビゲーションしてみないと検索目的に合致したページあるいはサイトであるかを判断できないことが原因と考える。検索結果に現れたページを起点としてサイト内をナビゲーションすることを考えると、ナビゲーションに役立つページを優先的に出力すべきである。逆に、ナビゲーションの助けにならないページはむやみに出力する必要がないと言うことになる。

つまり、サイトのホームページなどは、キーワード検索にヒットしなかったとしても積極的に出力すべきであり、逆にヒットしたページでも、PowerPoint スライドのような無駄なページは出力しないようにすべきである。

3. サイテーション・エンジンの機能

前節の要件に基づいて、サイテーションエンジンが採用した方式を述べる。

3.1. ページランク解析

ページランクは、そのページが WEB 全体においてどの程度の重要度を持つかどうかの指標である[Brin97,Google]。本方式のポイントは、

- 他のページからどのくらい参照されているかを重要度の指標とする。すなわち、他のページから多く参照されているページは、重要度が高くなる。
- 重要度は、そのページにたどり着く確率として表される。したがって、ページへ到達する確率をリンク数で割ったものが、リンク先のページへたどり着く確率として加算される。

また、図 1は本モデルの概念図である。

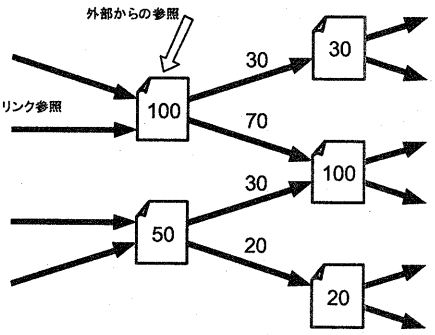


図1 ページランク計算のモデル

したがって、ページ u におけるページランク $R(u)$は、次式のように定義される。

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

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、$E(u)$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。

上記の式において、ベクトル R を固有ベクトルとみなして解くことで、、全ページに対するページランク値を得ることができる。

なお、ここで $E(u)$としたページアクセスの確率は、WWW のリンクから得るだけではなく、WWW サーバのアクセス数、メールマガジンや電子ニュースでの引用といった実世界からの引用数も利用できる。

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・ユニットは、同一テーマで編集された情報の範囲と構造をあらわす概念である。その算出のために、

- サイトとは、ドメイン/ホスト/ディレクトリのいずれかのレベル以下が同一であるページで構成される。
- サイトには、その顔となる代表ページが存在する。
- サイト内での情報は、代表ページをルートとする木構造をなしている。

という前提条件を考慮している(図 2参照)。これは、サイト作成者は通常、発信したい情報を同一ホストのあるディレクトリ以下に整理し、またホームページを起点として読み進めていけるように木

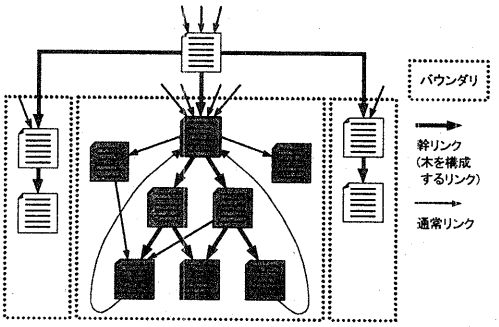


図2. インフォメーション・ユニット

構造を基本としたハイパーテキスト構成をとるものと考えるためである。

インフォメーション・ユニット判定アルゴリズムの概要は次のようになる。

代表ページの発見

Step 1: 同一ホストのページをグループ化

Step 2: 上記グループ内ページのうち、ホスト外部からのリンク参照数が最大のものを、代表ページとし、そのリンク参照数を L1 とする。

Step 3: 上記代表ページの傘下に存在し、ホスト外部からのリンク参照数がしきい値以上(通常 1)のものを選択する、選択されたページ数を N として、リンク参照数が L1/N よりも大きいものは、さらに代表ページとする。

Step 4: /index.html などのファイル名を持つものを代表ページとする。

バウンダリの発見

Step 1: 代表ページが属する最大のディレクトリを、サイトのバウンダリとする。

Step 2: 上記バウンダリ内に、サブバウンダリを発見した場合には、その部分を上位のバウンダリから差し引くこと。

情報木構造の発見

代表ページを起点とする幅優先探索をおこない、木構造とみなす。各ページには、代表ページからのリンク深さを付与する。

3.3. リンク構造ソート

リンク構造ソートは、キーワード検索によって選られたページ集合に対する一種のソートだが、下記に示すアルゴリズムの通り、通常のソートと異なる。すなわち、

Step 1: ページ集合をサイトごとにグループ化して、グループごとに代表ページを起点とする、木構造にしたがって、ソートする。

Step 2: この際に、ページ集合に含まれていない(キーワードを含まない)代表ページなどは、ページ集合に追加する

Step 3: ページ集合に含まれていても、木構造の下位に位置するなど、ユーザに提示する意味が低いページは削除する。

4. インプリメンテーション

サイテーションエンジンを構成する下記四つのコンポーネントのインプリメンテーションについて説明する。

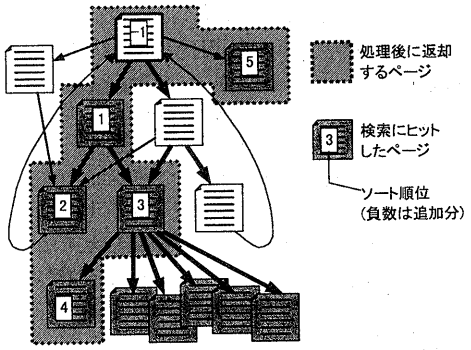


図 3. リンク構造要約の例

リンク DB	URL 情報、リンク情報を格納・管理
ローダ	リンク情報を解析・登録
アナライザ	ページランク解析、インフォメーション・ユニット解析を実施
フロントエンド	リンク構造ソートを実施

4.1. リンク DB

リンク DB は、サイテーション・エンジンが解析対象とする各種参照・引用情報を格納するデータベースである。

特に、リンク解析処理で頻繁に使用する、

- URL の情報(属性値、重要度、他)を得る。
- URL のリンク先 URL を得る。
- URL をリンクしている URL を得る。
- URL のフレーム内 URL を得る。
- URL が属するサイトを得る。
- サイトに属する URL を得る。

といった処理を高速化できるように設計した。データ管理には、フリーの B+Tree ライブラリである BerkeleyDB[BDB]を使用しているが、キーと値のペアに対するインデックス管理機能しか提供しないので、複数のインデックスを保持することで対処している。

また、一千万件を超える WWW ページならびにリンク情報を扱えるように、データベースファイルの分割管理を可能としている。

4.2. ローダ

ローダは、各解析処理で利用する外部情報をリンク DB に登録するコンポーネントである。

外部情報としては、

- Web クローラが出力するリンク情報ログ
- WWW サーバのアクセスログ

● 電子ニュースやメーリングリストの解析ログ

などを想定しており、現時点ではリンク情報ログだけに対応している。

4.3. アナライザ

アナライザは、ページランク解析ならびにインフォメーションユニット解析を実行するコマンドからなるコンポーネントである。各コマンドのインプリメントの詳細は本稿の範囲を超えるので割愛するが、概要を以下に述べる。

4.3.1. ページランク解析

ページランク解析における計算処理は、3.1節で述べた計算式を解くための、隣接行列および初期アクセスベクトルとページランクベクトルの掛け算の繰り返しである。

初期アクセスベクトルならびにページランクベクトルは、計算対象となるURL数=N個の要素を持ち、また隣接行列はN^2個の要素を持たなければならない。隣接行列はスパースなので、N^2個の要素を持つ代わりに、必要な要素だけを保持するデータ構造を用意した。いずれのデータ構造も、Nが100万件のオーダになるとマシン環境によってはメモリ上に確保するのが難しいため、一時領域に格納して必要な分を読み出して計算している。

初期アクセスベクトルは、現在はリンク参照情報に基づいて算出している。

4.3.2. インフォメーション・ユニット解析

本解析処理は、サーバ単位/サイト単位で実行される処理なので、サーバ単位で必要なデータをメモリ上に読み出し、3.2節で述べた解析処理を適用する.

4.4. フロントエンド

リンク構造ソート機能を提供する。本機能のインプリメント詳細は割愛するが、できるかぎり高速化の必要がある。今回の実装では、

- 同じ深さのページを取り出す
- 子ページを取り出す
- 親ページを取り出す

ことを念頭において、ページをあらわすデータは、それぞれ親・子・兄弟へのポインタをひとつずつ保持するようなデータ構造を持たせた。

5. 評価と議論

本章では、プロトタイプシステムを例にとって、サイテーション・エンジンの初期評価を行い、その結果を踏まえて議論する。

5.1. プロトタイプ・システムの構成

図 4に示すように、Web クローラシステム[Takano99]、ならびに全文検索システム[Akami96]に、サイテーションエンジンをアドオンする形で構成している。

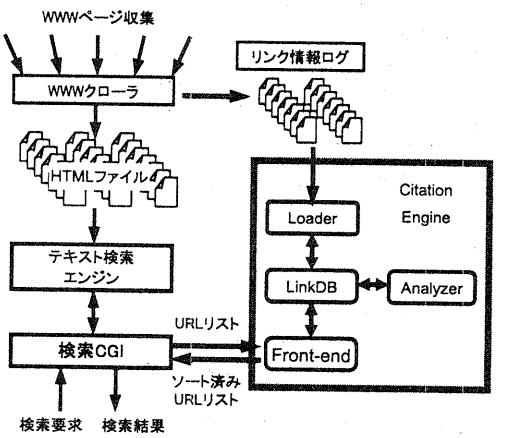


図 4 プロトタイプシステムの構成

5.2. 各機能の評価

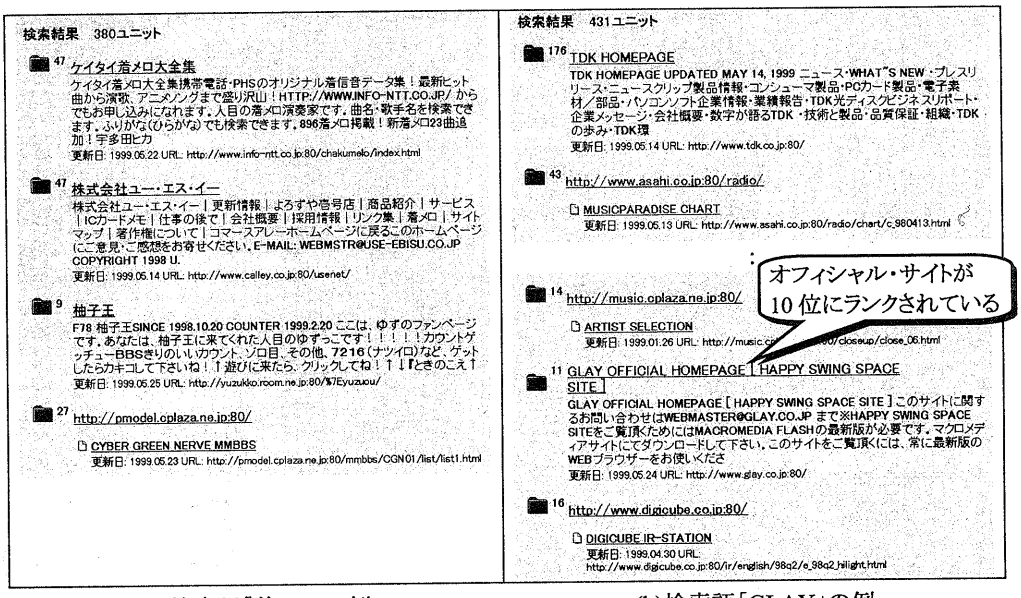
本節では、サイテーション・エンジンの各機能の評価について述べる。

5.2.1. 検索結果表示

サイテーション・エンジンの効果をはっきりとさせるために、検索結果の表示方式には次のような特徴を持たせた。

- サイト単位でクラスタリングし、これをユニットと呼んでいる。
- ヒット件数は、ページ数ではなくてユニット数を提示する。
- サイトの代表ページだけを提示し、タイトル脇のフォルダアイコンをクリックするとサイト内のヒットページのリストが現れる。
- 検索キーワードにヒットしていないホームページは、そのURLだけを表示する。

検索キーワードにヒットしたページで、ホームページからのリンク深さが2のものはタイトルと書誌事項を表示する。以下、ランキングシステムの初期評価について述べる。



(a)検索語「着メロ」の例　　(b)検索語「GLAY」の例

図5 サイテーション・エンジンを用いた検索結果の例

5.2.2. ページランクについて

まず、現時点では定量的な評価を行っていないので、使用した上での感想にとどめておくが、検索結果の一例を図 5に示す。

- たいていの場合、検索語にもっともふさわしいと思われるサイトが、上位に現れる。
- コンテンツとしての適合度を用いていないため、検索語を含んでいる人気サイトが上位に来てしまうことがある。特に、検索エンジンやプロバイダーサイトなどは外部からの参照数が多いために重要度が高くなり、また数々の情報を含むために検索結果の上位に現れてくることが多い。

5.2.3. 代表ページの判定について

他サーバからのリンク参照に基づいた代表ページ判定は、有効に機能している。

図 6は、代表ページ判定によって代表ページと判定されたURL のリストを抜粋したものであるが、企業サイトのようにサイトとサーバが等しい場合だけでなく、企業サイトの製品情報ページや、ISP サーバ内の個人サイトの判定がなされていることがわかる。

このため、代表ページを起点としたバウンダリ判定も有効に機能することとなり、サイトによるクラスタリングも妥当なものとなっている。

http://www.pc98.nec.co.jp:80/ http://www.pc98.nec.co.jp:80// http://www.pc98.nec.co.jp:80/Product/mg/ http://www.pc98.nec.co.jp:80/product/ http://www.pc98.nec.co.jp:80/product/MG/ http://www.pc98.nec.co.jp:80/service/ http://www.pc98.nec.co.jp:80/service/pccs/
http://www.nec.co.jp:80/ http://www.nec.co.jp:80/ad2000/ http://www.nec.co.jp:80/japanese/product/ http://www.nec.co.jp:80/japanese/whats-new/

(a) 企業サイトの例

http://www2n.biglobe.ne.jp:80/ http://www2n.biglobe.ne.jp:80/%7EWWESBJ/ http://www2n.biglobe.ne.jp:80/%7Ebakauma/ http://www2n.biglobe.ne.jp:80/%7Ef-kyoro/ http://www2n.biglobe.ne.jp:80/%7Egaku/ http://www2n.biglobe.ne.jp:80/%7Egala/mcast/ http://www2n.biglobe.ne.jp:80/%7Eju-/ : :

(b) ISP の例

図6 代表ページ判定の例

5.2.4. 情報構造ソートについて

情報木の判定は比較的うまく機能しており、ホームページやインデックスページの挿入、不要ページの削除など、検索結果のブラウジングしやすい表示を可能としている。

図 7は、検索結果表示の一部を抜粋したもの

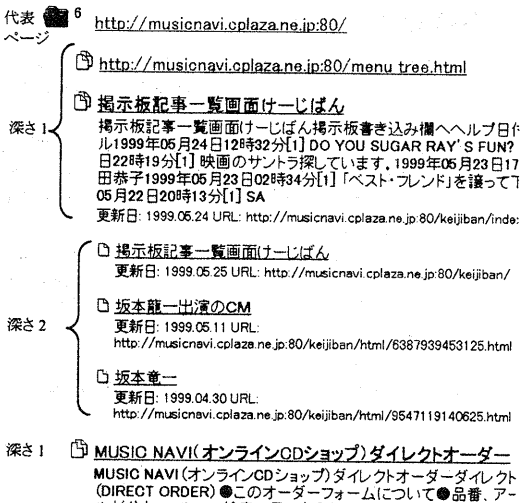


図7 情報木構造をもちいた表示例

であるが、情報木の構造にしたがってレイアウトされていることがわかる。

また、代表ページの挿入については、図5および図7において、ユニットのタイトルがURLになっているものが混じっている点に注意されたい。これは、検索語がヒットしなかった代表ページであることを示している。

しかし、情報木判定アルゴリズムは、WWWページ間の関係が木構造に近いことを前提としているため、完全グラフに近いようなWWWページに対してはうまく機能しない。いわゆるポータルサイトに多く、サイトマップと呼ばれる他ページへの目次が全ページに付与されているものが典型的な例である。こうしたWWWページに対しては、共通部分を取り除いて判定するといったヒューリスティクスが必要であるが、今後の課題である。

5.3. データベース評価

本節では、サイテーションエンジンの、データベース性能の評価結果を述べる。

なお、本評価を行った環境は、SUN Ultra5 (UltraSPARC-IIi 270MHz、メインメモリ 514MB、ハードディスク 9GB×2)である。また、登録したデータは、5/24~5/28 の期間、Web クローラが収集したデータを用いた。

5.3.1. データベース処理コマンドの実行速度

ある 100 万 URL のリンク情報の登録に、09:31:13 という結果が出ている。

5.3.2. IUA 処理の実行速度

IUA 処理は、ホームページ発見、バウンダリ設定、情報木判定からなる。LinkDB に 100 万 URL 登録されたある状態で、処理に 00:42:36(約 42 分)かかっている。データ規模から考えると、実サービスに適用するのに充分な速度であるといえる。

5.3.3. PRA 処理について

PRA 処理は繰り返し計算のため、収束状況ならびにかかる計算時間を評価した。前述の LinkDB において、06:30:34(約 6 時間半)かかっている。また、固有ベクトルが収束するまでの繰り返し回数は、30 回程度であった。

5.3.4. LSS 処理の性能

URL 数の上限を 500 件として実際の検索結果から生成した URL リストを 2000 件用意し、これを連続処理させた。その結果、最短で 0.1 秒、最大で 6.2 秒、平均で 0.2 秒の処理時間がかかっている。処理時間は、データベース検索用のインデックスがメモリキャッシュに載っているかどうかに依存しているため、必要なデータをメモリ上に展開することで、速度の向上を図れる。

5.4. 今後の予定

5.4.1. 他の引用情報の利用

3.1節でも述べたように、ユーザがある WWW ページを見に行く確率は、単にリンクの参照関係とは限らない。この観点から、外部の引用ソースを導入できるように機能拡張作業中である。

まず、WWW サーバのアクセス履歴を引用情報として加味し、その後、メーリングリスト・電子ニュースメッセージの引用情報の加味を検討する。

5.4.2. アンカー文字列の活用

5.2.2節で述べたように、ページランクを用いたランキングはリンク参照関係だけを用いているため、検索語によっては必ずしもベストのサイトがトップにくるわけではない。これは、検索語がページに含まれているかどうかだけを調べていて、検索語とページの適合度を考慮していない点が問題である。しかし、検索語とWWWページの適合度として従来の統計的手法がかならずしも適当とは言えない。

このため、[Chakra98]でも採用している「リンク元ページに含まれているアンカー文字列と検索語の適合度を考慮する」ことで、ページランクに内容への適合度を加味する方式を検討している。

6. 関連研究

サイテーション・エンジンのベースとなったのは、Hub/Authority ページの算出[Kleinb98]を代表とするIBM Almaden 研究所の CLEVER プロジェクト[Clever]、ならびにグローバルなリンク解析を用いたページランクの算出を特徴とする Stanford 大学の Google サーチエンジン[Google, Brin98]である。

一方、局所的なリンク構造解析を行うことで WWW ナビゲーションを支援するツールを提供するというアプローチも以前から行われている。たとえば、[Mukhej97]ではサイト内の WWW 構造を可視化するツールを提供している。また、[Takano98]ではユーザのナビゲーション履歴からページ重要度を判別して、自動的にブックマークを生成する機能を提供している。

WWW サイトの構造を解析して検索結果の整理をするアプローチは、いままでも広く試みられているが、階層構造を推定しようとするものはまだ少ない。[Harada99]は、主にディレクトリ構造とファイル名を用いて階層構造を推定して効果を得ているが、本論文とは異なりサイトとそのホームページの判定は実施していない。また、ランキングもテキスト統計処理に基づくものである。

7. おわりに

リンク解析を用いた WWW 検索ランキングシステムであるサイテーションエンジンについて、開発の背景、ランキングシステムが採用したアルゴリズム、インプリメンテーション、プロトタイプシステムを用いた初期評価、ならびに今後の研究開発予定について述べた。

WWW 検索システムは、インターネットサービスだけではなく、企業内のナレッジマネジメント・ツールとしてきわめて重要である。サイテーション・エンジンは、従来の情報検索技術とは異なるアプローチによって有効な検索ランキング機能を提供できる。

8. 謝辞

日頃よりご討論いただく、NEC ヒューマンメディア研究所、古関義幸部長、神場知成マネージャ、杉浦主任、ならびに高島洋典部長、福島俊一主任研究員、赤峯亨主任、山田洋志主任、松田勝志主任に感謝いたします。

また、開発を担当した、新潟NESの津谷主任、鈴木氏に感謝いたします。

参考文献

[Akami96] "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

[Brin98] Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," pp107-117, Proceedings of 7th WWW Conference, May 1998

[Chakr98] Soumen Chakrabarti, et al., "Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text," Proceedings of 7th WWW Conference, May 1998

[Fuku99] 福島, 松田, 高野 "Web ページの重要度ファクタに関する一考察,"情報知識学会研究発表会, May 22, 1999

[Harada99] 原田, 佐藤, 風間, "WWW ページ間の階層構造の推定と検索システムへの応用," pp105-112, 情処研報 DBS-118-14, May 1999

[Kleinb98] John Kleinberg. "Authoritative sources in a hyperlinked environment," Proceedings of 9th ACM-SIAM Symposium on Discrete Algorithms, 1998

[Lawren99] Steve Lawrence and C. Lee Giles, "Accessibility of information on the web," pp107-109, NATURE, Vol. 400, 8 July 1999

[Mukhej97] Sougata Mukherjea and Yoshinori Hara, "Focus+Context Views of World-Wide Web Nodes," pp.187-196, ACM Hypertext'97

[Takano98] Hajime Takano and Terry Winograd, "Dynamic Bookmarks for WWW", pp297-298, ACM Hypertext'98

[Takano99] Hajime Takano and Nobuya Kubo, "Development of a Scalable Web Crawler," pp337-339, NEC R&D, Vol.40, No.3, July 1999

URL

[BDB] http://www.sleepycat.com/

[Clever] http://www.almaden.ibm.com/cs/k53/clever.html

[Google] http://www.google.com/