4 WWW情報検索技術と評価の問題

福島 俊一　NEC 情報通信メディア研究本部インターネットシステム研究所
t-fukushima@cj.jp.nec.com

インターネット上のWWWコンテンツは，論文・新聞・特許のような文献資料と比べ，検索対象としてみたときの性格に大きな違いがある．日々変化する大規模なハイパーメディアである点，コンテンツのタイプや質が多様な点などである．そのため，WWWコンテンツの自動収集技術，大規模な並列検索システム技術，検索語頻度のほか新鮮度・引用度・人気度・ページタイプにも着目した情報価値判断技術など，WWW情報検索に特徴的な新技術の研究開発が進んでいる．しかし，現状のテストコレクションでは，このようなWWW情報検索技術を部分的にしか評価することができず，総合的評価を可能にするようなWWW用テストコレクションの構築・整備が望まれる．

■WWWサーチエンジンの技術■

あらゆる情報がインターネットで提供されるようになりつつある．十数年前であれば，情報検索システムというと，図書館の蔵書目録を検索するシステム，新聞記事や特許公報のオンラインデータベース，辞書CD-ROMの検索システムなどが思い浮かべられたであろうが，今日，一般ユーザに最も身近な情報検索システムは，インターネット上のWWW（World Wide Web）コンテンツを対象とした情報検索システム，いわゆるWWWサーチエンジンであろう．インターネットへアクセスする際の入口となるポータルサイトの多くが，WWWサーチエンジンをその中核に持っている．

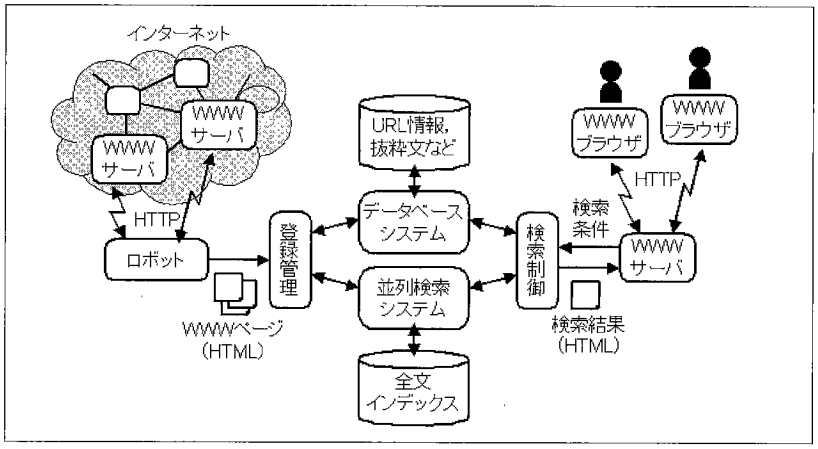
WWWコンテンツは，論文・新聞・特許のような文献資料と比べ，検索対象としてみたときの性格に大きな違いがある．WWWは，全世界に張りめぐらされたクモの巣に例えられるような，大規模なハイパーメディアである．インターネット上には10億ページ近い膨大なWWWコンテンツが存在する．日々どこかで，新しいコンテンツの作成，古いコンテンツの更新が行われ，WWWは変化している．ニュース記事・製品カタログ・リンク集などさまざまなタイプのコンテンツ，また，公式ホームページから個人メモ風のものまでさまざまな質のコンテンツが混在している．

WWWサーチエンジンでは，このような性格を持つWWWコンテンツを対象とするために，WWW情報検索に特徴的な新技術の研究開発が進められている．具体的には，WWWサーチエンジンの処理の流れ（**図-1**参照）に沿って，以下のような技術があげられる．

第1はWWWコンテンツの自動収集技術である．インターネット上に分散して存在するWWWコンテンツを，ロボット（あるいはクローラ，スパイダー）と呼ばれるプログラムで自動収集する．ロボットはWWWのハイパーリンクを次々にたどってコンテンツを集めていくが，すべてを巡回するのには相当の時間がかかる．高い情報鮮度を確保するために，複数サイトの並列収集，更新頻度の高いサイトや特定トピック・重要コンテンツの選択的収集など，収集の高速化・効率化が図られている．

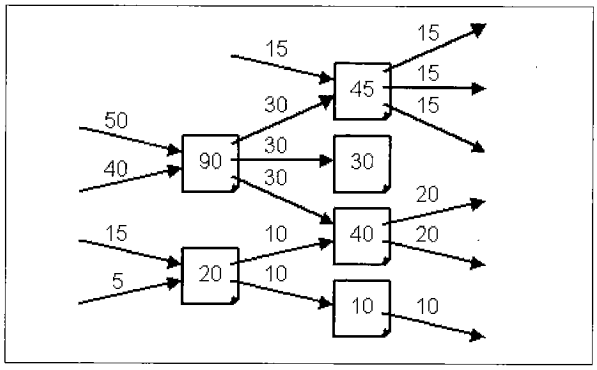
第2は大規模な並列検索システム技術である．WWWコンテンツの規模は，たとえば500万ページ分のテキストでおよそ25ギガバイトにもなる．また，アクセスが集中する時間帯であれば，1秒間に数十から数百という検索回数をこなすことが必要になる．このような高い要求性能・スケーラビリティを満たすために，大規模なインデックスを複数の計算機に分割して検索する手法（テキスト規模増大に対する検索レスポンスの改善），検索回数を複製システムで効率よく分担してこなす手法（同時アクセスに対するスループット向上）などの並列検索技術が不可欠になっている．

第3はWWWコンテンツの価値判断技術である．WWWサーチエンジンのユーザが入力する検索語は，多くの場合，1語ないしは2語である．これに対して，数百から数万ものWWWコンテンツがヒットし，実際にユーザがコンテンツ内容を見るのは最初の数件から十数件程度であることが多い．そのため，ユーザの欲しいコンテンツが上位10件程度の中に含まれるように，WWWコンテンツの価値を判断する手法が重要になる．この詳細は次章で述べる．



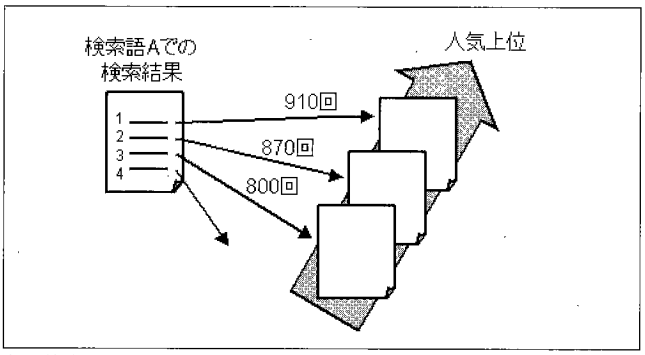
ロボットによって，インターネット上のWWWページが自動収集される．収集されたWWWページは全文インデックスに登録されると同時に，そのURL情報・抜粋文などがデータベースに登録される．検索時には，検索条件が全文インデックスと照合され，ヒットしたページについて，データベースから検索結果表示に必要な情報が読み出される．

図-1　WWWサーチエンジンにおける処理の流れ



各WWWページの引用度スコアは，そのページが他ページから受けているハイパーリンクの重みを累積することで計算する．これによって，多数引用されるほどスコアが高くなる．さらに，各々のハイパーリンクの重みとして，リンク元のページの引用度スコアに応じた値を用いる．これによって，重要なページに引用されるとスコアが高くなる．

図-2　WWWページの引用度スコアの計算方法



ある検索語と，その検索結果からユーザ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分かる．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．

図-3　WWWページの人気度に基づくランキング方法

■WWWコンテンツの価値判断■

以下では，WWWを構成する各ページの重要度を判定するためのファクタとして，代表的なもの5つを紹介する．WWWサーチエンジンでは，このようなファクタのいずれか（1種あるいは複数種類）が検索結果の絞り込みや順位付けに用いられている．

①検索語頻度：「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．TF・IDFスコアがよく知られている[1]．TF値（Term Frequency）は，検索語が多数出現するページほど大きくなる．その際，IDF値（Inverted Document Frequency）を掛け合わせることで，ありふれた検索語の重みを小さく抑える．WWWコンテンツを対象とした場合，HTMLタグを考慮し，タイトルや見出し部分に検索語が出現すればスコアに加点するというような改良も加えられている．

②新鮮度：「新しいページほど重要」という考えに基づき，ページ更新日時の新しいものを優先する．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，情報の価値や量が高まっている可能性が高い．

③引用度：「多数引用されるページは重要」「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．WWWにおいては，ページAからページBへのハイパーリンクが引用関係（ページAがページBを引用している）を表すので，リンク構造を解析することで引用度を計算することができる（図-2参照）[2), 3)]．

④人気度：「多数のユーザが参照したページほど重要」という考えに基づく．人気度は過去の参照履歴から計算できる（図-3参照）．引用度・人気度ともWWWコンテンツの評判のよさを表す指標であるが，引用度はWWW構造から求める静的かつグローバルな評判であり，人気度は特定サーチエンジンの参照履歴から求める動的かつローカルな評判だといえる．

⑤ページタイプ：「ユーザの問題解決タスクに合ったタイプのページほど重要」という考えに基づく．たとえば，パソコン購入という問題解決タスクには「製品カタログ」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．このようなページタイプに着目して，検索結果の絞り込みを実現する（図-4参照）．WWWページの各々がどのようなページタイプに当てはまるかは，WWWページの構造的な特徴（たとえば，HTMLタグやハイパーリンクの使い方，URLの文字列，画像の有無，各タイプに特徴的なキーワードなど）に着目することで自動判定できる[4]．

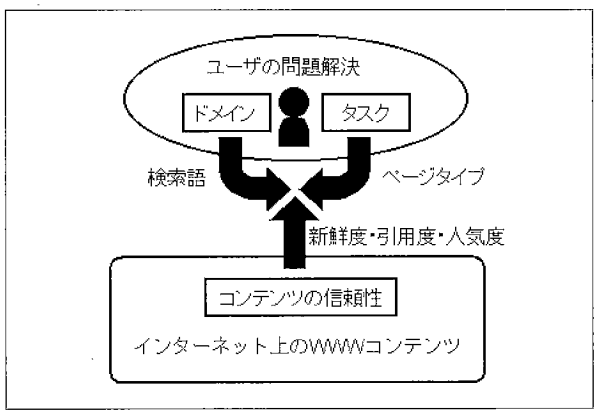
論文・新聞・特許などの文献資料を対象にするならば，最初に述べた検索語頻度が中心的なファクタになる．このファクタに関して，数多くの研究・改良が重ねられてきている．しかし，WWWサーチエンジンのユーザが入力する検索語は1・2語程度にすぎず，一方，WWWコンテンツの作成者は，自分のページを多くの人の目に触れさせたいから，WWWサーチエンジンにヒットしやすい言葉を

NETPLAZA（http://netplaza.biglobe.ne.jp/）では，ページタイプサーチの機能があり，検索語を入力するほかに，「製品カタログ」「リンク集」「掲示板・チャット」「プレゼント情報」「調査報告」「求人情報」「イベント情報」などのページタイプを指定することで，検索結果を絞り込むことができる．

図-4　ページタイプに着目したWWW情報検索

作為的にページに埋め込むことさえするという．このような問題に対して，検索語頻度のみでは価値判断のファクタとして限界があるため，ほかのファクタも組み合わせることが必要になってきた．

なお，上であげた5つのファクタの間には，次のような関係があると考えられる（図-5参照）．WWWサーチエンジンはユーザの問題解決手段ととらえられる．このとき，検索語は主に問題解決ドメインを絞り込み，ページタイプは主に問題解決タスクを絞り込む役割を果たす．一方，残る3つのファクタ（新鮮度，人気度，引用度）は，ユーザの個々の問題解決とは独立に，WWWコンテンツの信頼性を計るための尺度だと考えられる．



検索語は問題解決のドメインを絞り込み，ページタイプは問題解決のタスクを絞り込む．一方，新鮮度・引用度・人気度はWWWコンテンツの信頼性を計る尺度である．

図-5　WWWコンテンツの重要度ファクタ間の関係

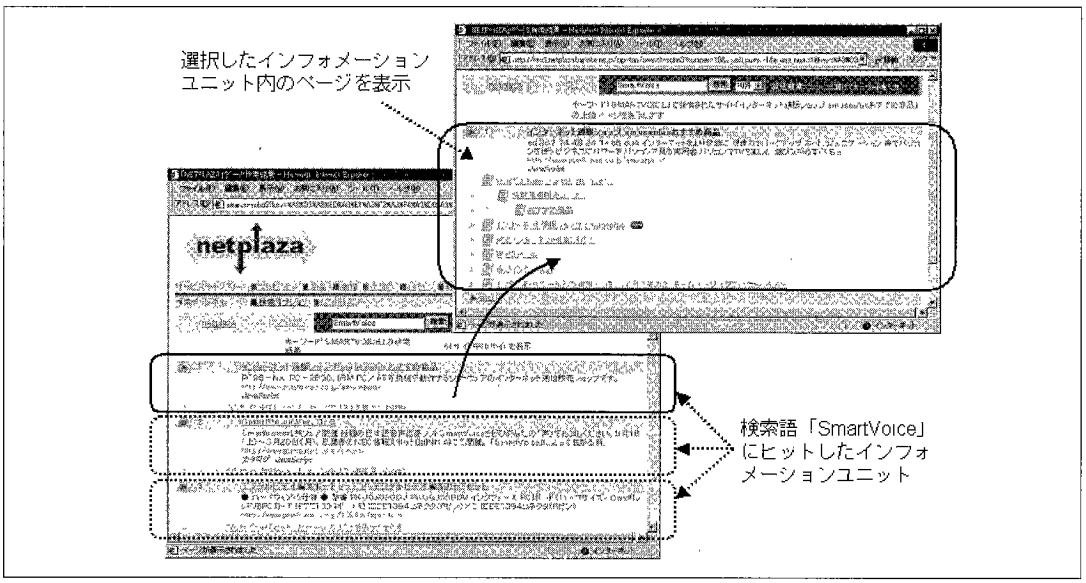
■WWW情報検索技術の評価にかかわる問題■

テストコレクションは情報検索技術の評価・改良に不可欠なものとして，これまでの情報検索研究の発展に大きな貢献を果たしてきた．しかし，WWW情報検索技術の評価のために，現状のテストコレクションは必ずしも十分なものではない．以下，テストコレクションを構成する文書集合・検索課題・正解判定という3側面から各々課題をあげる．

文書集合の問題：まず，実際のWWWコンテンツを一度に大量に集める必要がある．コンテンツが日々変化するため，段階的に増やすのではリンク構造などに不整合が生じてしまう．また，すべてのWWWを網羅することは不可能であるから，収集範囲をどう限定するのが妥当かについても考えねばならない．さらに，本文以外にもコンテンツの価値判断に必要な各種情報（更新日時，リンク構造，参照履歴，HTMLタグ情報など）が併せて提供されないと，前章で述べたような重要度判定手法が適用できない．

検索課題の問題：論文・特許などを対象とした場合とWWWを対象とした場合とでは，ユーザの検索要求は傾向が異なるであろう．実際のWWW検索の利用場面に即した検索課題を用意する必要がある．

正解判定の問題：従来のテストコレクションでは，文書は各々独立したものとして正解判定されている．しかし，WWWコンテンツは複数ページで1つのトピックを表すように構成されていることが多い．正解ページそのものにヒットしなくとも，正解ページへのリンクが得られれば十分有用である．検索結果をページ単位に列挙するのではなく，複数ページからなるインフォメーションユニット（あるいはサイト）単位に提示するWWWサーチエンジンも現れている（図-6参照）[3]．また，WWW内にはミラーサイトのようにまったく同じ内容が重複して存在することがあり，正解判定時に考慮が必要になる．さらに，精度評価の



NETPLAZA（http://netplaza.biglobe.ne.jp/）では，サイトファインダーの機能を使うと，検索結果がインフォメーションユニット単位にまとめて表示される．

図-6　インフォメーションユニットに基づくWWW情報検索

指標についても，WWW情報検索の利用場面に即したものになるよう，議論していく必要があろう．

以上で述べた問題は，WWW情報検索の精度，すなわち情報価値判断技術の評価にかかわる問題である．しかし，現実のWWWサーチエンジンの改良に結び付くような評価を目指すならば，前述したようなWWWコンテンツ自動収集技術や大規模な並列検索システム技術を切り離して考えるわけにはいかない．すべてのWWWを網羅することは現実的に不可能なので，ロボットによる収集範囲・収集方法をどう決定するかが，WWWサーチエンジンの検索精度・性能を大きく左右する．また，きわめて高速な検索レスポンスを維持できることが，方式実装の前提となる．このような側面も併せて考えたWWW情報検索評価の方法論を確立していくことが必要になる．

ここで述べたWWW情報検索技術の評価およびWWW用テストコレクションの問題に対して，TREC-8（1999年のText Retrieval Conference，米国商務省国立標準・技術院NISTが主催）では，サブトラックとしてWWW情報検索が取り上げられた．Webトラックという名称で，100ギガバイト規模のWWWコンテンツによる大規模Webタスクと，2ギガバイト規模のサブセットを用いた小規模Webタスクが実施された[5]．小規模Webタスクでは，WWWページ本文だけでなくページ間リンク情報も付加された文書集合が与えられ，リンク構造を利用した検索技術の効果についても評価が試みられた．大規模Webタスクでは，処理速度・容量，ハードウェア資源，スケーラビリティなどの面からの比較もなされた．しかし，小規模Webタスクにおいてリンク構造を利用することの効果がほとんどみられないという結果になり，WWWコンテンツ規模が小さいために十分なリンク情報を含んでいないのではないか，WWW情報検索に適した検索課題や評価方法を用いるべきではないか（TREC-8のWebトラックでは従来TRECの検索課題や評価方法をそのまま用いた），などの疑問があがった．TREC-9では，小規模Webタスクを10ギガバイト規模に拡大し，検索課題も実際のWWW検索のログをもとにして作成するなど，改良が加えられるとのことである．Webトラックは，まだスタートしたばかりの試みであり，本章で述べたような問題の多くが課題として残されているのが現状である．

今後，TREC Webトラックの回が重ねられていけば，WWW用テストコレクションとして，より改良・整備されたものが構築されていくものと期待できる．国内でも，日本語テストコレクションの構築・拡充に向けた意欲的活動が進められており，同様の期待が寄せられる．しかし，その一方で，従来型のテストコレクションという枠組みの延長では，WWW情報検索のある一面しか評価できないのではないかという限界も感じる．テストコレクションという枠組みでWWW情報検索のどんな側面を評価するか，さらに，テストコレクションを超えるWWW情報検索評価の方法論が考えられないか，について議論を深めていきたい．

参考文献
1) Salton, G. and McGill, M. J.: Introduction to Modern Information Retrieval, McGraw-Hill (1983).
2) Brin, S. and Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine, Proceedings of 7th WWW Conference, pp.107-117 (1998).
3) 高野　元, 久保信也: 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, 情報処理学会研究報告, DBS-120-2 (2000).
4) 松田勝志, 福島俊一: 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, 情報処理学会研究報告, FI-53-2 (1999).
5) Hawking, D., Voorheers, E., Craswell, N. and Bailey, P.: Overview of the TREC-8 Web Track, Proceedings of TREC-8 (1999).
(平成12年6月30日受付)