WWW情報共有のための検索機能の開発＊

6L－1

荻久保 友史　栗田 雅芳 †

株式会社 東芝　東京システムセンター ‡

1. はじめに

近年のインターネットにおけるWWWの急速な普及により、企業から個人まで、幅広くWebページを作成し、情報発信を行うことが可能になってきている。また、WWWを企業内情報システムとして活用するイントラネットも急増してきている。従来、IS部門などの特定部門のみが可能であった情報発信を、部門毎にWWWサーバを構築することも珍しくなくなってきた。そのような状況のもとにおいては、WWWからの情報発信量は加速度的に増加していくことになる。必要とする情報がどこにあるのかわからず、探すのには非常に苦労する。

そのため、ある特定サイトのWWWサーバ情報を高速に検索したいという要求が発生してきてた。

2. WWW情報の検索

WWW情報の検索システムでは、通常、指定されたWWWサイトの情報に対応する検索インデックスを生成し、そのインデックスをもとにWWW情報の検索を高速に行う。このような検索システムでは、次の二点について検討が必要である。

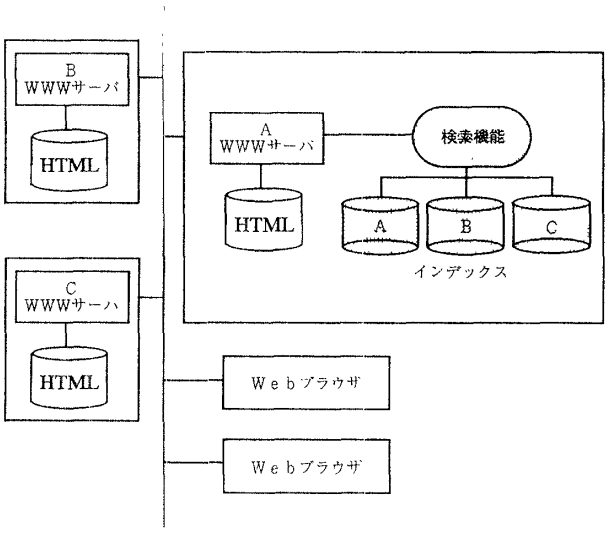

図1　WWW情報の検索のしくみ

(1) 検索インデックスの更新

WWWサイトの情報は日々更新され続けている。そのため、対応する検索インデックスもそれに合わせて更新していかなければならない。

(2) 検索キーワードの設定

WWWからの情報を検索するため、あらかじめキーワードとなる言葉の検索システムへの登録設定が必要であるとすると、そのためのメインテナンスを考慮しなければならない。

3. 検索機能の概要

今回開発を行ったWWW情報検索機能の全体の概要を図1に示す。検索機能は、自分の場所のWWWサーバの情報も含め、指定したWWWサイトの情報をあらかじめ収集し、それらに対応した検索インデックスを生成する。情報の検索は、この検索インデックスをもとに行われる。2章で述べた検討項目について、次に示す対応を行っている。

(1) 検索インデックスの自動更新

今回開発した検索機能には、検索インデックスを自動的に更新するためのWWWサイト情報自動収集ロボット機能を備えている。インデックスメインテナンスのための人手をかけることなく、常に新しいWWW情報検索環境を提供することができる。

(2) WWW情報の全文検索

今回の検索機能の検索エンジンでは、対象とするWWWサイトのHTML情報を全文検索することができる。あらかじめキーワードを登録するといったわずらわしい作業は必要なく、また利用者は、思い付いた単語で自由に検索を行うことができる。

4. WWW情報の検索機能

今回開発したWWW情報の検索機能部分の構成を図2に示す。検索機能は大きく分けて、インデックス管理部、インデックス更新部、検索実行部から構成されている。それぞれについて説明する。

---

* Development of WWW Information Retrieval System

† Tomofumi Ogikubo, Masayoshi Kurita

‡ Toshiba Corporation Tokyo System Center

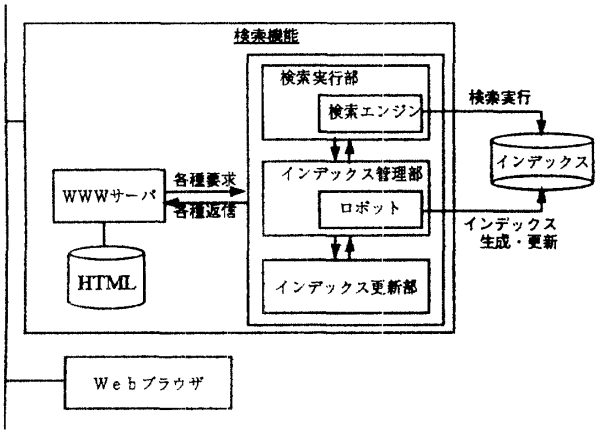

図2　WWW情報検索機能の構成

4.1　インデックス管理部

インデックス管理部では、検索インデックスの初期生成を含めたインデックス管理、自動収集ロボットを起動しての WWW サイトの情報収集・インデックス更新、インデックス更新部や検索実行部とのやりとりなど、検索機能全体の管理を行っている。

4.2　インデックス更新部

インデックス更新部は、自動収集ロボットの WWW 検索対象サイト先訪問スケジュールの設定処理や、その設定したスケジュールに従って自動収集ロボットの実行要求をインデックス管理部に対して行っている。

4.3　検索実行部

検索実行部は、利用者からの検索要求に基づいて WWW 情報検索を実際に行う。利用者からの検索要求はWebブラウザを通して入力され、インデックス管理部に入る。その後インデックス管理部からの指示により検索実行部は検索処理を実行する。検索処理は、検索エンジンによって高速に行われ、検索結果が利用者に返される。

5. 検索機能の構築

WWW 情報検索環境の構築例として、今回開発した WWW 情報検索機能を利用し、サンプルとして構築した10,000ページから成るWWWサイトを検索対象とする検索環境の構築について紹介する。

5.1　構築環境

WWW 情報検索環境の構築は、東芝 UNIX ワークステーション AS4080 、OS は Solaris 2.4 、メモリ 64MB 、WWW サーバとしては、Netscape Communications Server を使用して行った。

5.2　検索対象サイトの指定

検索対象サイトの設定は、Web ブラウザより検索機能の設定画面を通して検索対象サイトの URL を入力するのみである。また、その指定した WWW サイトからどれだけリンクをたどった先までを検索対象とするかの指定も同時に行う。リンク先を深く設定すればそれだけ得られる情報も多いが、その分、必要となるインデックスのサイズも大きくなってしまう。今回の検索機能では、インデックスサイズは、対象となるWWWサイトHTMLテキストデータ量の 1.5 ～ 2 倍程度となる。

5.3　更新スケジュールの設定

インデックスの生成・更新の処理の実行中は、その対象となっている WWW サイトの情報検索は行うことはできない。検索対象となる WWW サイトのページ数が増えてくると、インデックスの更新にも時間がかかるようになる。そのため、インデックス更新スケジュールは注意して決めなければならない。例えば、毎週日曜日や、深夜に更新するようなスケジュール設定が必要となる。

5.4　WWW 情報検索の実行

今回の検索機能においては、検索処理は、検索エンジンにより検索インデックスを利用して高速に行われる。内部的には 1 秒程度の時間で検索処理は終了する。Webブラウザを通して検索結果を表示するため、ネットワークの混雑状況にもよるが、実際に結果を得るまでにはもう少し時間がかかることになる。検索結果は、検索キーワードに対してヒットした Web ページの URL 、タイトル、最終更新日付等が Web ブラウザ上に表示される。

6. おわりに

今回の WWW 情報検索機能は、既存の WWW サーバに対して特別なハードウェアも必要なく、容易に検索機能を付加することができるようにするものである。世の中のWWW情報を検索するために、Yahoo! 、InfoSeek をはじめとする多くの WWW 検索サービスが存在する一方で、イントラネットとして WWW を利用したい場合など、ある特定の WWW サーバサイトに対する情報検索機能を構築したいという要求がある。そのような場合には、今回開発を行った WWW 情報検索機能が有効になると考えられる。