WWWページ指向の解析機能を持つ検索サーバ

3P−10

西村英樹† 稗田薫 吉川耕平

シャープ株式会社マルチメディア推進本部システム開発センター

1. はじめに

WWWページの爆発的増加により，様々なWWW検索エンジンが登場している[1),2)]．当社もホームページ上で検索システム「SMART ENGINE」を提供している．SMART ENGINEを含め多くの検索エンジンはロボット型と呼ばれるもので，ロボットプログラムが自動的にHTMLで記述されたWWWページを収集し，索引作成を行う．ロボット型の検索エンジンでは，自動化することを目的としているため，索引(インデックス)の作成も画一的な方法がとられている．一般的な方法としては，タイトルを見出しより重く，見出しを本文より重くして索引を作成する[3)]．そのため，タイトルに組織名称やファイル名，作成日付などが入っている場合には，検索精度が著しく低下する．また，最近ではこのような検索システムでのヒット率を向上させるために意図的に異なるキーワードを入れている例もある．

これらの問題を確実に解決する方法として人手によりページ内のキーワードを抽出して索引を作る方法があるが，莫大な時間を必要とする．本稿では，出来るだけ少ない作業でこのような問題を解決する手法について述べる．

2. 階層ごとのパーザ

一般的にWWWサイトが異なったり，同一のサイト内においてもディレクトリ階層が異なるとWWWページのレイアウトポリシーは異なることに着目し，各サイトまたは各階層毎に専用パーザを開発し検索精度の向上を目指す．しかし，対象のサイト数に比例して，開発工程がかかるため，共通化できる部分を徹底的に共通化する必要がある．

† A Web Search System with Page Oriented Parsing
Hideki Nishimura, Kaoru Hieda, Kouhei Yoshikawa
Sharp Corporation, Corporate Research and Development Group

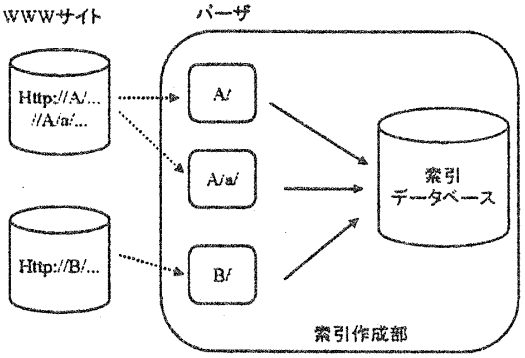

図1 パーザ

パーザの要求仕様は次のようなものがある．1) タイトル，本文を設定可能にする：一般にWWW検索サーバでは，タイトルの部分と本文には異なる重みで扱うので，これらの部分を自由に設定できる必要がある．2)1つのパーザ開発の労力が小さい：専用パーザの最適化処理としてタイトルを特別に設定するだけで十分な場合も多い．デフォルトの解析処理との差分だけを数行程度の設定ファイルに記述できれば開発の労力が小さく理想である．3) 複数のパーザを統一的に扱える：パーザを開発する毎に検索サーバ本体に修正を加えることのない仕組みが必要である．4) 複数の階層が同じパーザを共有する．そこで，パーザの適用範囲も階層構造を持たせ，例えばあるURLドメインに対し，原則としてパーザAを適用するが，例外的にディレクトリ/b/以下の階層に対してはパーザBを適用する，それ以外のドメインはパーザCを適用する，とする．図1に例を示す．http://A/に対しては原則としてパーザ“A/”を適用し，例外的に/a/以下はパーザ“A/a/”を適用する．http://B/に対してはパーザ“B/”を適用する．

3. 実装と評価

前節で述べた機能を持つWWW検索サーバをWindows NT上に試作した．図2に構成を示す．一般のロボット型WWW検索エンジンと同じ構成をとり，

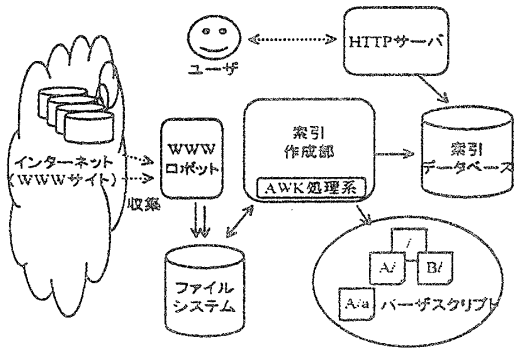

図 2　システム構成図

WWW ロボット, 索引作成部, HTTP サーバからなる. 索引作成部には簡便かつ柔軟にプログラムを記述できる AWK の処理系を利用し, AWK のスクリプトでパーザを記述できるようにした. パーザではタイトル, URL, キーワード, 更新日時を登録する.

まず WWW ロボットが WWW ページを収集し, ファイルシステム上に保存する. 収集完了後, 索引作成部はファイルシステム上の各 HTML ファイル毎に対応するパーザスクリプトを選択し索引作成する. パーザ "/" は他の該当するパーザがなかった時のためにデフォルトのパーザとして用意しておく. 索引作成後, HTTP サーバはユーザからのアクセスのたびに索引を参照し, 検索結果を HTML の形式で出力する.

評価を次の要領で行った. 1) 全ての WWW ページを同一のパーザ A で作成した索引 I_A と, あるサイトのディレクトリ以下の WWW ページをパーザ B で作成, それ以外をパーザ A で作成した索引 I_{AB} を用意する. 2) 索引 I_A, I_{AB} をあるキーワード K で検索する. 3) キーワード K でヒットすべきページ P_K のヒット件数 H_A, H_{AB} とヒット順位 R_A, R_{AB} を比較する.

評価対象を, http://www.sharp.co.jp/とし, パーザ A を/以下のデフォルトパーザ, パーザ B を /sc/excite/cook/text/以下のページのタイトルを最適にするパーザ (図 3), K={"メニュー","レンジ"}, P_K=「液晶ナビゲーションレンジ "メニューさん"」とした結果を表 1 に示す. /sc/excite/cook/text/以下のレシピページは, タイトルが全て「今日のメニュー」となっており, 各ページの内容を適切に表現しているものではなかったが, パーザの開発により適切なタイトルが設定され, 目的のページが多量の検索結果に埋もれることなく検索できることが分かった. なお, I_{AB} でヒット件数が 653 であるのは検索方法が OR 検索であり, キーワード「レンジ」を含むレシピページが多かったためであるが, 重みが小さく検索結果の上位には来なかった.

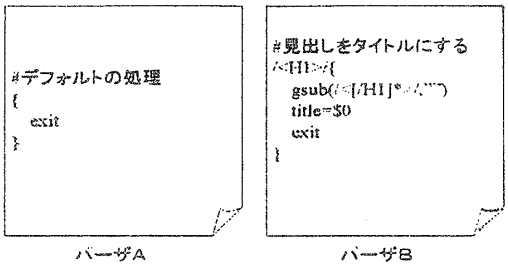

図 3　パーザスクリプト

表 1　ヒット件数および順位

	I_A	I_{AB}
ヒット件数	722	653
ヒット順位	183	6

4. おわりに

本稿では URL 階層に従ってパーザを選択し, 最適な索引作成を行う WWW 検索サーバを試作, 評価した. 単純なタイトルの設定や, 見ための統一感を出すための共通ヘッダやフッタを取り除き, 意味のある内容だけを扱うことで検索精度が向上した. 専用パーザの開発工程の削減を目指し, 複数パーザを統一的に扱う枠組にスクリプトを用いることでコンパイル, 組み込みなどの時間が短縮でき, C 言語などでの開発に比べパーザ作成時間を大幅に短縮できた. また, 階層毎にパーザを設定することで開発するパーザ数を減らすことが出来た. 本システムの他の応用として, 1 ファイルあたりのキーワード数を重み計算に用いているシステムでは, アーカイブされたものを仮想的に分けるパーザを開発することによって, 検索精度を向上させることが出来る. また, テキスト形式のファイルでもメール (メーリングリスト) などのフォーマット化された文書ならば, 専用のパーザを開発することで HTML と同じようにキーワードに重み付けして扱え検索精度が向上すると考えられる.

参考文献

1) 林良彦, 小橋喜嗣: WWW 上の検索サービスの技術動向, 情報処理学会論文誌, Vol. 39, No. 9, pp. 861–865 (1998).
2) 西村英樹, 河野浩之, 長谷川利治: WWW データ資源検索システムの実装と評価, 情処研報 96-DBS-109, Vol. 96, No. 68, pp. 263–268 (1996).
3) 西村英樹, 河野浩之, 長谷川利治: 知識領域分割による検索キー提示機能の評価, アドバンストデータベースシンポジウム ADBS'98, pp.79–86 (1998).