WWW上のカスタマイズ可能な検索システムに関する研究

3J－5

伊藤英二 山本晋一郎 濱口毅 阿草清滋
名古屋大学工学研究科地圏環境工学専攻

1 はじめに

現存するWWW上での情報検索システムのほとんどは，HTMLテキストをロボットにより取得して，加工したものに対してキーワードによる検索を行うものである．これらのシステムは文字列及びそれに論理式を加えた質問文でしか検索を行うことができない．これらのシステムがこのような単純な検索しかできない理由の一つは，膨大なHTMLテキスト群を対象としているために，そこから検索用データを抽出する場合には検索用データをなるべく小さくする必要がある為である．また，これら大規模な検索システムの他に，例えばある組織に関する情報が調べたいときに，その組織ローカルなWWW検索システムがあるとひじょうに便利だが，検索システムが用意されているところは少ない．本研究の目的は，比較的中規模から小規模なものを検索対象として考え，様々な条件を含んだよりきめこまやかな検索条件による検索ができる検索システムを構築し，ユーザが求めている情報を発見する作業を支援することである．

本稿では，現存する検索システムにはない検索条件を提案し，次に検索システムの概要を説明し，最後に実現した内容について説明する．

2 検索条件

本来WWWとは，空間や時間を超越した知識共有を目指したものである．しかし，現存する検索システムでは，WWW上の情報全てを同列の物として扱っており，サーバから取得したデータ毎の性質による区別は特になく，また，そのデータによって作られた検索システムで検索時の条件として使えるものは，キーワードによるマッチング及び，それにAND,ORなどの論理式を加えた程度のものである．しかし，WWW上のテキストは

- 最終更新時間
- 更新頻度
- テキストサイズ
- そのテキストに張りつけられているグラフィックスの数
- 書かれている言語
- そこから張られているリンク数
- 被リンク数

など，そのテキスト自身の性質を表すいろいろな情報を含んでいる．これらのうちいくつかが検索条件として使うことができるなら，ユーザにとってより的確な検索ができるようになるはずである．例えば，論文を捜す場合なら，ある程度の量があるテキストを捜した方がいい，ここ1週間以内に更新されたテキストという情報も最新の話題を追っていく場合などには有益である．

A Customizable search engine for WWW
Eiji Ito,Shinichirou Yamamoto,Takeshi Hamaguchi,Kiyoshi Agusa
Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-01 Japan

3 検索システムの概要

本研究で製作する検索システムにおいては，よくあるWWWの検索システムと同様にWWWロボットと呼ばれるプログラムによってデータを取得する．取得したデータをインデクシングし，同時に付加的な情報もデータベースに登録し，検索できるるようにする．

3.1 ロボットについて

まず，データが何も無い状態では幅優先探索によりテキストを取得する．これは，幅優先探索の方が特定のサーバにアクセスが集中して負担が重くなることを避けることができると思われるからである．次回からデータ取得も，基本的には幅優先探索だが，前回のデータ取得で得られた更新頻度の情報を使って，更新頻度が高いと思われるページを優先的に探索する．

3.2 更新頻度情報の利用

WWW上には様々な更新頻度を持つテキストがあるが，この性質はデータ取得にも反映されなければならない．すなわち，更新が緩慢なテキストを頻繁に取りにいってもネットワーク上の余計なトラフィックを産むだけであるし，更新が頻繁なテキストは頻繁にデータを取得しなければ，すぐに古くなってしまう．ここで，更新頻度の情報はデータ取得時に更新時間を記録し，その値から計算することによって得られる．サーバからテキストの最終更新時間が返ってくる場合は，その値から更新されてからどれだけ経っているかを計算し，更新頻度の値を上下させる．

あるページのテキストに動的なテキスト更新メカニズムが使われている場合やサーバプログラムによっては，テキストの最終更新時間を返してくれないので，最終更新時間が明らかで無い場合も多い．このような場合は，テキスト自体のサイズとテキストから生成したチェックサムの値を記録しておいて，前回のデータ取得時からの変更があるか無いかを調べて，更新頻度情報の値を変化させる．

3.3 検索システムのインターフェイス

検索システムのインターフェイスとしてはキーワードを入力する検索を基本として，必要に応じて条件を付加する形で行う．具体的には，サイズに関しては，[考慮しない][0KB ～ 100KB][100KB ～ 1MB][1MB ～] などを，更新時間に関しては，[考慮しない][1日以内][1週間以内][1ヶ月以内][1年以内][1年以上] などが選択できるフォームを用意する．追加の条件を選択しなかった場合は単なるキーワードの検索として機能する．

4 検索システムの試作

現在，上に述べた検索条件のうち今回は，最終更新時間，更新頻度，テキストサイズが使用可能な検索システムを作成している．WWWロボットのプログラム自体はスクリプト言語Perlによって書いた．データのインデクシング及び検索実行部分には，ffwというC++で書かれた英語の文章をインデクシングし検索することのできるパッケージを，形態素解析ツールJUMANを使うことにより日本語の文章も処理できるようにしたものを用いた．

5 おわりに

まだ，プロトタイプしか実現できていない．早く実現して，実際に運用することで，ユーザから指摘された問題点などを考察し，この検索システムのアプローチによる有効性を考えていきたい．また，条件が多くなり，操作体系が複雑になってくると使いづらいものになってしまう．複雑な条件を使うことができて，しかも初心者にも優しい検索システムを考えていかなければならない．