D-6

専門検索サイトの動的統合による次世代検索システムDAISENにおける検索サイトエディタの開発

A Wrapper Editor for Search Sites

山田泰寛* 廣川佐千男†
Yasuhiro Yamada Sachio Hirokawa

1. はじめに

爆発的に増え続けるWWW上の情報をどのようにすれば活用できるかという問題は，現在も将来も情報化社会における重要な問題である．我々は，この膨大な量の情報の中から必要な情報を探す際に，Yahoo!，Google等の一般の検索エンジンを使用するが，その検索結果の品質が大きな問題となっている．それに対し特定のテーマの情報に限定した専門検索サイトが増えている．

検索者の意図する分野に特化した検索サイト群が見つかり，あたかも一つの検索エンジンのように機能するメタサーチエンジンが実現できれば，WWW上の情報検索はより効率のよいものになる．専門検索サイトを用いてメタサーチエンジンを構築する際には，どの検索サイトを選択するかが問題となる．後藤等 [2] は，シソーラスを利用することによる検索サイトの選択手法を提案している．我々は，検索者の目的に応じた専門検索サイトを動的に選択し，一括した統合検索を行なうシステムDAISEN [1]を開発している．

WWW上に多数存在する専門検索サイトを統合し検索を行なうためには，その検索サイトに検索を行なうために必要となる情報や，異なる検索サイトの検索結果を統合するために検索結果から必要な部分を抽出する情報など，専門検索サイトごとに情報管理を行なう必要がある．DAISENにおいて，このような検索サイト情報は自動生成され，その情報は高い精度を持つが，間違いがないとは限らない．そこで，実際の検索サイトのページを参照しながら，管理者が検索サイト情報を効率良く編集するためのツールが必要である．また，WWW上のWebページと同様に検索サイトも更新される．多数の検索サイト情報をデータベースに格納していても，サイトが更新され，情報が古くなってしまえば使い物にならない．よって，システムの保守の観点からも，人間が検索サイト情報の編集を行なうツールが必要である．

本研究ではこのような要求から，検索サイトエディタを開発した．検索サイトエディタは検索サイト情報を管理者がGUIから編集するためのツールである．このようなツールは見方を変えると，半構造化データから情報抽出を行なうためのツールであるといえる．半構造化データから情報の抽出を行なうプログラムは，一般にラッパーと呼ばれる．

ラッパーの自動生成の研究については [4, 8] 等多くあるが，実用的観点からは生成されたラッパーの検証はさらに重要である．Web上の異なる情報源の統合を目指すTSIMMISプロジェクト [1] においては，Webページのフォーマットの変化や，Webサイトのアップデートの問題から，生成されたラッパーをGUIから編集するためのツールの必要性が述べられている．他にJantke等 [3] も確認のためのGUIシステムを実装し実験している．Webデータのフォーマットは頻繁に変更されるので，一度生成したラッパーがいつまでも正しいとは限らない．Kushmerick等 [5, 6] は，このような問題から，ラッパーのメンテナンスは重要な問題であると述べている．

2. DAISENにおける検索サイトエディタの位置付け

DAISENは大きく分けて3つの機能によって構成される．1つ目は検索サイト情報の自動生成機能を持つ「データ加工部」である．2つ目は検索サイトとディレクトリ構造の情報を格納するための「データ蓄積部」である．3つ目は複数の検索サイトに検索キーワードを与えて得られる結果を統合し，ユーザーに提示する「統合検索部」である．検索サイトエディタは「データ加工部」に含まれる．

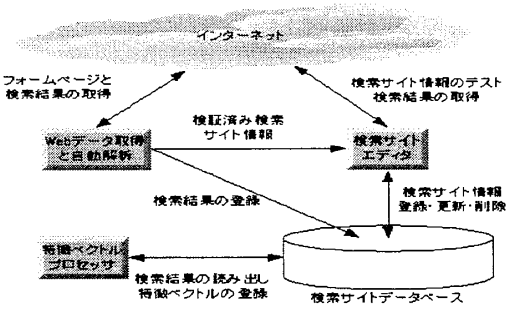

図 1: データ加工部

データ加工部は「Webデータ取得と自動解析」,「検索サイトエディタ」,「特徴ベクトルプロセッサ」の3つの項目によって構成される (図 1).

DAISENにおいて新しく使用される検索サイトについては，Webからのデータ取得と自動解析により，検索サイトエディタで読み込み可能な検索サイト情報が生成される．検索サイトエディタでは，必要があればそれを修正して実際にその検索サイトに対し検索が行なえるかどうかチェックして，検索サイトデータベースに登録する．この他に検

*九州大学大学院システム情報科学府
†九州大学情報基盤センター

[1]Directory Architecture for Integrated Search Engine, http://daisen.cc.kyushu-u.ac.jp

索サイトエディタはデータベースから既に登録されている検索サイト情報を読み出して，更新，削除等の保守管理操作の機能も持つ．

3. 検索サイトエディタの機能

検索サイトエディタには，検索サイトに検索をかけるために必要な情報の解析と編集，ラッパー作成と編集，検索実行による検索サイト情報のテストなどの機能を持つ．ここで，検索サイト情報とは，検索サイトの URL，検索を行なう際に必要となる検索フォームのパラメータ，検索結果1件を表すタグパターン (ラッパー)，自動解析部で作成されたラッパーの精度を示すスコア [7] などを指す．

検索サイトエディタでは，自動解析部で作成された検索サイト情報の含まれるファイルを選択することにより編集を行なう．編集を行なった検索サイト情報はデータベースに保存される．

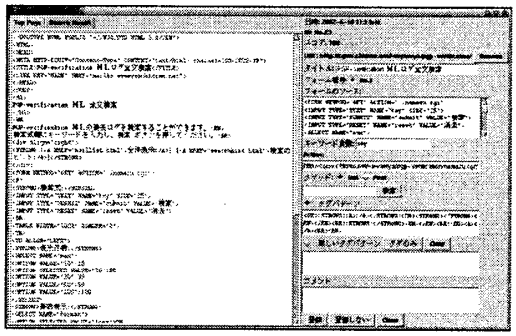

図 2: 編集画面

検索サイトエディタを起動し，自動検証部で作成された検索サイト情報を読み込むと，スコア，検索サイトの通し番号，検索サイトのタイトル，URL が表示される．編集ボタンを押すと，図 2 の編集画面でそのサイトの情報の修正ができる．

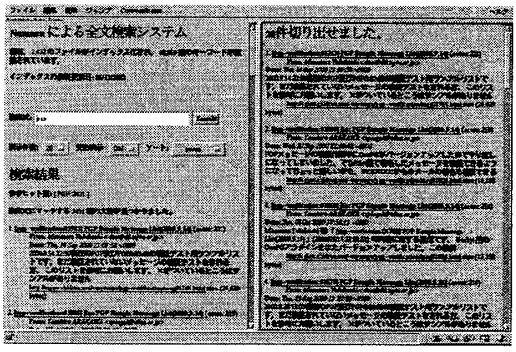

図 3: ラッパーの確認画面

図 2 の画面左側には，編集中の検索サイトのトップページのソースが表示され，必要な部分をコピー&ペーストできる．画面右側にはフォーム情報とタグパターンなどの情報が表示され，編集が可能になる．検索サイト情報が正しいことを確認するために，編集画面内の検索フォームで検索を実行すると，ブラウザが立ち上がり，画面左側には検索サイトから返された検索結果が，右側にはラッパーによって抽出した結果が表示される (図 3)．編集者はこの 2 つを見比べて検索サイト情報が正しいかどうかを確認する．

以上の操作を経て，編集を行なった検索サイト情報をデータベースに登録する．

また，初期画面で検索サイトの URL を指定することにより，その検索サイト情報をデータベースから読み出す．そして，その情報を編集画面に表示し，更新，削除するといった保守管理操作も行なうことができる．

4. まとめ

本研究では，我々が開発した検索サイトエディタについて紹介した．検索サイトエディタは，自動生成された検索サイト情報を GUI で編集するためのツールである．また，データベースから検索サイト情報を読み出して，更新，削除等の保守管理操作の機能も持つ．この検索サイトエディタは，DAISEN において実際に使用されている．

今後の課題としては，自動検証部分を検索サイトエディタの中の機能として含めることや，データベースに存在する検索サイト情報の一覧を表示する機能の追加などがあげられる．

参考

本研究は，情報処理振興事業協会 (IPA) の委託により財団法人ソフトウェア工学研究財団 (RISE) が実施した平成13年度「高度情報化支援ソフトウェアシーズ育成事業」による成果であり，また，株式会社ヒューマンテクノシステムによる支援による．

参考文献

[1] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman and J. Widom. The TSIMMIS Project: Integration of Heterogeneous Information Sources, Proc. of IPSJ Conference, pp. 7–18, 1994.

[2] 後藤将志, 大圃忠親, 新谷虎松, 選択的メタサーチエンジンにおけるシソーラスを用いたサーチエンジン選択手法の提案, 人工知能学会論文誌 17(3), pp. 285–292, 2002.

[3] G. Grieser, K . P. Jantke, S. Lange and B. Thomas, A Unifying Approach to HTML Wrapper Representation and Learning, Springer LNCS 1967, pp. 50–64, 2000.

[4] N. Kushmerick, D. S. Weld and R. B. Doorenbos, Wrapper Induction for Information Extraction, IJCAI97, pp. 729–737, 1997.

[5] N. Kushmerick, Regression testing for wrapper maintenance, AAAI-99, 1999.

[6] N. Kushmerick, Wrapper Verification, World Wide Web Journal, 3(2), pp. 79–94, 2000.

[7] 酒井美由紀, 廣川佐千男, 検索サイトラッパー検証のための検索結果件数推定方法, 第 13 回データ工学ワークショップ, 2002.

[8] Y. Yamada, D. Ikeda and S. Hirokawa, Automatic Wrapper Generation for Multilingual Web Resources, (投稿中).