九州大学学術情報リポジトリ
Kyushu University Institutional Repository

統合検索システムDAISEnでの検索サイトフォーム分析

山田, 泰寛
九州大学大学院システム情報科学府

松永, 吉広
九州大学大学院システム情報科学府

野口, 正人
九州大学大学院システム情報科学府

中藤, 哲也
九州大学情報基盤センター

他

https://hdl.handle.net/2324/2981

出版情報：情報処理学会研究報告 ： データベースシステム. 2003 (71), pp.49-54, 2003-07. 情報処理学会
バージョン：
権利関係：ここに掲載した著作物の利用に関する注意 本著作物の著作権は（社）情報処理学会に帰属します。本著作物は著作権者である情報処理学会の許可のもとに掲載するものです。ご利用に当たっては「著作権法」ならびに「情報処理学会倫理綱領」に従うことをお願いいたします。

KYUSHU UNIVERSITY

社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

統合検索システムDAISEnでの検索サイトフォーム分析

山田　泰寛† 　松永　吉広† 　野口　正人† 　中藤　哲也†† 　廣川佐千男††

† 九州大学大学院システム情報科学府 〒812-8581 福岡県福岡市東区箱崎 6-10-1
†† 九州大学情報基盤センター 〒812-8581 福岡県福岡市東区箱崎 6-10-1
E-mail: †{yshiro,matsunaga,noguchi}@matu.cc.kyushu-u.ac.jp, ††{nakatoh,hirokawa}@cc.kyushu-u.ac.jp

あらまし　統合検索システム DAISEn は，WWW 上の専門的な検索サイトを動的に統合するメタサーチ自動生成システムである．多数の専門的検索サイトへの検索を統合することにより，膨大な WWW の情報源に対し効率良く高精度の検索を実現する．統合検索のためには，検索サイト毎に異なる入力形式に対応してクエリを送らなければならない．DAISEn では，検索ページの解析によりサイト毎の入力形式を獲得している．本論文では，国立国会図書館関西館データベース・ナビゲーション・サービス Dnavi に登録された 2882 件の検索サイトを対象として，フォームの解析及び入力形式の抽出を行った．

キーワード　メタサーチ，検索エンジン，ディープウェブ

Query Form Analysis of Search Sites in Meta-Search System DAISEn

Yasuhiro YAMADA†, Yoshihiro MATSUNAGA†, Masato NOGUCHI†, Tetsuya NAKATOH††, and Sachio HIROKAWA††

† Graduate School of Information Science and Electrical Engineering, Kyushu University 6-10-1 Hakozaki, Higasi-ku, Fukuoka 812-8581, Japan
†† Computing and Communications Center, Kyushu University 6-10-1 Hakozaki, Higasi-ku, Fukuoka 812-8581, Japan
E-mail: †{yshiro,matsunaga,noguchi}@matu.cc.kyushu-u.ac.jp, ††{nakatoh,hirokawa}@cc.kyushu-u.ac.jp

Abstract　DAISEn is a meta-search engine generation system which integrates specialized search engines dynamically. Meta-search engines, in general, require a program to deal with a different format of query for each site. DAISEn analyzes the query format of the search page and generates a wrapper automatically. As an empirical evaluation of DAISEn, the present paper evaluates the extraction of the query formats of 2882 search sites registered on NDL Database Navigation Service Dnavi.

Key words　Meta Search, Search Engine, Deep Web

1. はじめに

爆発的に増え続ける WWW からの情報検索技術は，現在も将来も情報化社会における重要な技術である．この膨大な量の情報の中から必要な情報を探すには，Yahoo!，Google 等の一般の検索エンジンは必須であるが，検索に該当した件数が膨大であったり，検索結果のランキングにおいて，探したいページが上位に来ないなど，検索結果の品質が大きな問題となっている．

一方，多くの企業などのサイトにおいて利用者へ直接情報を提供するため，自サイト内の情報やデータベースについて検索サービスを提供しているものが増えている．WWW 全体を対象とした検索エンジンと対比して，このようなサイトを**検索サイト**と呼ぶ．検索サイトが提供する情報は，データベースから動的に作成されるため，一般の検索エンジンではカバーされない．そのため，これらは Invisible Web [13], [14]，Deep Web [2] あるいは Hidden Web [4], [5] と呼ばれる．検索サイトは特定のテーマに限定した情報を提供しているため，その品質は高いと考えられる．検索者の意図する分野に特化した検索サイトに対して検索を行なえば，精度の高い検索を効率良く行なえる．

WWW に存在する複数の検索エンジンもしくは検索サイトを統合するシステムは**メタサーチエンジン**と呼ばれ，SavvySearch [3]，mamma [8]，vivisimo [16]，askOnce [1] はその代表である．メタサーチエンジンは複数の検索サイトを扱うため，

品質の良い結果が得られるように思われる．しかし，これらメタサーチエンジンの扱う検索サイトは固定されており，新たな検索サイトを動的に追加することはできない．

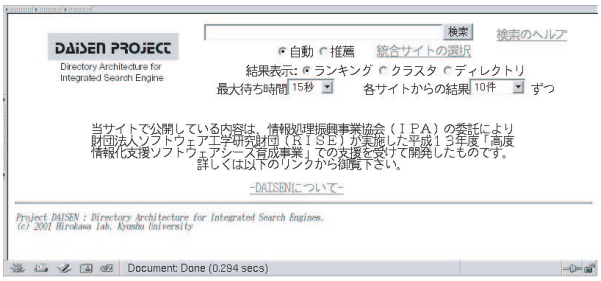

図 1　統合検索システム DAISEn

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** を開発している (図 1) [12], [15]．DAISEn では，ユーザが検索を行なう際に投げるクエリから関連性の高い検索サイトを動的に選びそれらを統合する (図 2)．

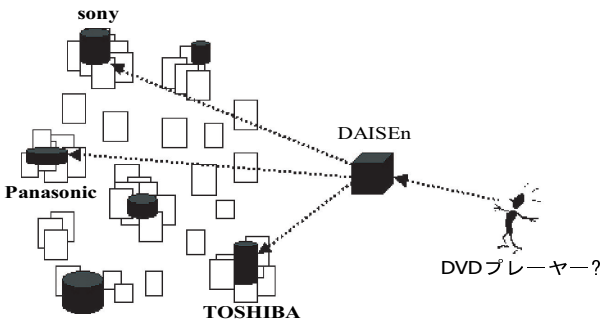

図 2　検索サイトの選択

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，各検索サイトへクエリを送るために必要となる入力形式 (**フォーム情報**) や，検索結果から必要な部分を抽出するためのパタンなど，専門検索サイトごとに情報管理を行なう必要がある．これらの情報をまとめて**検索サイト情報**と呼ぶ．これまで我々は，フォーム情報の自動生成法 [9], [10]，パタン抽出によるラッパーの自動生成法 [7]，ラッパーの精度の自動評価のための検索結果の推定手法 [11]，検索サイト情報の編集のための検索サイトエディタ [17] 等の開発を行なって来た．

本論文では，WWW 上のデータベースへのリンクを提供している国立国会図書館関西館データベース・ナビゲーション・サービス Dnavi [6] から収集した 2882 件の検索サイトに対して，フォームの解析を行ない，フォーム情報の抽出についての実験とその評価を行なった．

本論文は以下のように構成されている．2 節において，DAISEn の概要について述べる．3 節において，データベース・ナビゲーション・サービス Dnavi に登録されている検索サイトについて述べる．4 節において，収集した検索サイトのフォーム情報の自動抽出について述べる．最後に，まとめと今後の課題について述べる．

2. 統合検索システム DAISEn

DAISEn では，WWW 上に存在する検索サイトの情報をあらかじめデータベースに登録しておく．検索者から検索の要求があったとき，検索者の投げたクエリから関連の高い検索サイトをデータベースから自動的に選ぶ．そして，選ばれた検索サイトそれぞれに対して検索を行ない，得られた検索結果を統合し検索者に提供する．また，検索者はカテゴリ分けされた検索サイトから手動で対象とする検索サイトを選択することもできる．

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，ユーザに提示する「統合検索部」である．

データ加工部では，各検索サイトから検索結果を得るために必要となる入力形式 (**フォーム情報**) 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 (出力パタン情報) を作成する．この 2 つの情報を合わせて**検索サイト情報**と呼ぶ．

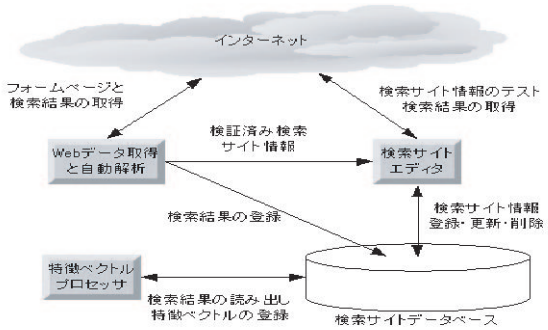

図 3　データ加工部

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 つの機能によって構成される (図 3)．DAISEn に新しく追加される検索サイトについては，WWW 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 [17] 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人手でチェックして，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. データベース・ナビゲーション・サービス Dnavi

本節では，DAISEn における検索サイト情報の自動生成の評価のために使用する国立国会図書館関西館データベース・ナビゲーション・サービス Dnavi (図 4) から集めた検索サイトにつ

図 4　データベース・ナビゲーション・サービス “Dnavi”

いてのフォームの特徴などについて述べる．

Dnavi では，WWW 上に存在するデータベースへのリンク並びにデータベース検索サービスを提供している．ただし，Dnavi は本論文で述べるようなメタサーチを提供するものではない．収録データベース数は 2003 年 6 月 10 日現在約 7,000 件である．このようなデータベースの中には，検索機能を提供している検索サイトが数多く存在する．

3.1　検索サイトの収集と解析

まず，Dnavi の検索システムに “検索” というクエリを用い検索を行なうことでキーワード検索ができるデータベースを選んだ．これは検索サイトならば，そのページ中に “検索” という単語が現れると考えたためである．次に，得られた 2931 件のサイトから人手で

（1） 検索サイトである
（2） リンクを辿れば検索サイトに辿り着けそうである
（3） 検索サイトではない

の 3 種類に分類した．その結果，(1) にあたるものが 1846 件存在した．(2) にあたるものが 624 件存在した．(2) にあたるページの例として，複数の検索サイトへのリンクがはられているリンク集のページが多数存在した．そこから実際にリンクをたどることにより 1036 件の検索サイトを得た．こうして合計 2882 件の検索サイトを得た．

これらの検索サイトのトップページを収集し，テキストボックスの数に着目して解析を行なった (表 1)．テキストボックスとは，クエリを記入するフォームのことを指す．収集した検索サイトのうち，単一のテキストボックスを用いている検索サイトは 559 件，全体の約 19%であった．複数のテキストボックスを用いる検索サイトは 1531 件存在し，全体の約 53%であった．

複数のテキストボックスを用いる検索サイトとして，これらを用いて AND 検索や OR 検索が行なえる検索サイトと，それぞれのテキストボックスに属性が割り当てられている検索サイトの主に 2 種類が存在した．後者の例として，図書館のサイトにおいて，著者，本名などの属性を持つ複数のテキストボックスを用いている検索サイトがあった．また，単一のテキストボックスを用いている検索サイトでもプルダウンメニューやチェックボックスと組み合わせているサイトが数多く存在した．

表 1　テキストボックスの数

テキストボックスの数	件数	テキストボックスの数	件数
0 個	793	14 個	91
1 個	559	15 個	13
2 個	167	16 個	6
3 個	236	17 個	2
4 個	249	18 個	18
5 個	162	19 個	1
6 個	126	20 個	1
7 個	89	21 個	6
8 個	84	23 個	2
9 個	47	24 個	1
10 個	41	26 個	1
11 個	56	74 個	1
12 個	79	100 個	1
13 個	50		

表 2　テキストボックスが 0 個の検索サイト

フレーム形式	547
他のページへ移動	8
テキストボックスをもたない検索サイト	64
解析不能	174

テキストボックスの存在しない検索サイトは 793 件存在し，全体の約 28%であった (表 2)．この内，検索サイトのトップページがフレーム形式となっているページが 547 件存在した．これらのページには，フレームのいずれかにテキストボックスなどが存在したが，トップページを収集したため，テキストボックスがカウントされなかった．また，テキストフォームは存在せずラジオボタンやプルダウンメニューで構成される検索サイトが 64 件存在した．

解析不能に含まれるものとして，検索サイトの URL に個人情報や時間などが含まれるものがあった．この URL はその時にのみ有効な URL であり，後からその URL を使ってトップページを収集しようとしてもエラーを表示するページが返って来た．

以上より，全体の内で 1 つのテキストボックスのみで検索を行なうサイトは少なく，複雑な検索を行なうサイトが多いことが分かった．また，トップページを収集するだけでは，収集できない検索サイトも存在することが分かった．

検索サイトのトップページがフレーム形式であったり他のページへ移動するものに対して，DAISEn はトップページの収集しか行なわないため，その検索サイトは統合検索に利用できない．しかし，このような検索サイトも数多く存在するため，これらも扱わなければならない．この問題は，各フレームのソースの解析を行なうことにより，容易に収集が行なえると考える．しかし，URL に個人情報や時間が含まれる検索サイトに対しては，現状では収集は困難である．

また DAISEn は，対象とする検索サイトとして，1 個のテキストボックスを提供しているサイトを想定している．2 個以

上のテキストボックスを提供している検索サイトに対しては，個々を独立したテキストボックスとして扱うため，例えば，複数のテキストボックスを用いて AND 検索，OR 検索を提供するサイトに対しては，AND 検索，OR 検索を行なうことはできない．テキストボックスが複数ある場合，デフォルトでは 1 番目のテキストボックスに対して検索を行なうが，他のテキストボックスを対象としたい場合は，検索サイトエディタにより手動で変更する．

4. フォーム情報の作成

DAISEn では検索サイトの URL を入力として受け取り，検索結果を得るために必要となるフォーム情報と検索結果から必要な部分を抽出するための出力パタンを生成するまでに以下の手順が必要となる．

（1） 検索サイトのトップページの収集
（2） フォーム情報の解析
（3） 検索結果の収集
（4） 出力パタンの生成

まず，検索サイトのトップページを収集し，その HTML ソースからフォームに関連するタグを解析することによりフォーム情報を作成する [10]．フォーム情報とは，CGI のメソッド (GET もしくは POST)，クエリを投げるために必要となる URL，キーワード変数から成り立つ．キーワード変数とはクエリを値に持つパラメータのことを指す．例えば，“Yahoo! JAPAN(注1)” において，フォームに関するタグは，以下のように記述されている．

```
<form action=http://search.yahoo.co.jp/bin/search>
<input size=30 name=p>
<input type=submit value=検索 >
</form>
```

この時，クエリを投げるために必要となる URL は “http://search.yahoo.co.jp/bin/search?” であり，キーワード変数は “p” である．実際に “網走” というクエリで検索をかけると，検索結果のページの URL は “http://search.yahoo.co.jp/bin/search?p=%CC%D6%C1%F6(注2)” となる．

次に，生成されたフォーム情報とあらかじめ決めておいたキーワードの集合から検索結果を取得し，出力パタンの生成を行なう．

本節では，Dnavi から収集した検索サイトを用いて，フォーム情報の生成について実験を行なう．

4.1 評　　価

3.1 節で収集した検索サイトのうち，1 つのテキストフォームを持つ検索サイト 559 件に対してフォーム情報の生成の実験を行なった．しかし，これらの内にはテキストボックスが存在するが，検索機能を提供していないサイトや，ユーザの ID を記入するなど特別なクエリを与えなければならない検索サイトがいくつか存在した．そこで，これらを省いて 501 件に対して実験を行なった．作成されたフォーム情報を用いて，URL を作成し検索結果が得られるかどうかで成否を判断した．

メソッドに関して，GET が 152 件，POST が 349 件存在した．表 3 は，それぞれのメソッドを持つ検索サイトとフォーム情報の抽出の成功率である．GET に関しては 87%，POST に関しては 76%の成功率であり，全体の成功率は 79%であった．

表 3　フォーム情報作成の成功率

メソッド	件数	成功件数 (成功率)
GET	152	132 (87%)
POST	349	264 (76%)

失敗する原因として，単純なフォーム情報の抽出漏れの他に，JavaScript を用いているためフォーム情報が抽出されない場合が存在した．また，HTML タグの使用の間違いなどのためにフォーム情報が抽出されない場合も存在した．このようなサイトに対するフォーム情報の抽出は現状では不可能である．

また，検索サイトのトップページからブラウザ上でクエリを投げなければ検索結果を得られないページが存在した．このようなサイトに対しては，フォーム情報は作成できるものの，検索結果にはエラーが返ってきた．このようなサイトに対しても現状では検索結果は得ることは不可能である．

以上のような現状ではフォーム情報の抽出及び検索結果の獲得が不可能なサイトを除けば，成功率は 8 割を越えた．また，単純なフォーム情報の抽出漏れを改善すれば，さらに成功率は良くなると期待できる．

他にフォーム情報の抽出に失敗した例として，テキストボックスにクエリを投げると同時に，チェックボックスやプルダウンメニューを選択しなければならないサイトが存在した．例えばクエリを投げると同時に対象とするデータベース等を選択しなければならないサイトが存在した．しかし，HTML ソースの中にどれを選択するか指定がない場合，現状ではいずれの選択も行なわないため，検索結果の獲得に失敗した．

実験では 1 つのテキストボックスを持つ検索サイトを対象としたが，同時にチェックボックスやプルダウンメニューからいくつか項目を選択できるものが存在した．DAISEn では，他の項目を選択したい場合は，検索サイトエディタにより手動で変更する．

今回の実験において，テキストボックスの存在する検索サイトでも，検索機能を提供していないサイトなどの様に統合検索には利用不可能な検索サイトが存在するため，検索機能を提供しているかどうか，DAISEn において利用可能な検索サイトであるかどうか，あらかじめ判断する機能が必要である．

DAISEn ではテキストボックスが複数ある場合，個々を独立したテキストボックスとして扱う．このため，テキストボックスを 1 つ含む検索サイトと同様に，それぞれのテキストボックスからフォーム情報の作成が可能である．よって，実験で使用した 2882 件の検索サイトの内，テキストボックスが存在する 2090 件の 8 割以上の検索サイトに関して，フォーム情報の作成が可能である．

(注1)：http://www.yahoo.co.jp/
(注2)：2 バイト文字は URL の中ではエンコードされる．つまり，%CC%D6%C1%F6 は “網走” をエンコードしたものである．

5. まとめ

本論文では，データベース・ナビゲーション・サービス Dnavi から収集した 2882 件の検索サイトを用いて，テキストボックスに着目したフォームの分析及びテキストボックスを 1 つ含む検索サイトに対してフォーム情報の作成の実験を行なった．この結果，統合検索に利用不可能な検索サイトを除けば，8 割以上の成功率でフォーム情報が抽出できることが分かった．

実験において，複数のテキストボックスを提供するサイトや，チェックボックスやプルダウンメニューと組み合わせるサイトなど，複雑な検索機能を提供するサイトが多く存在することが分かった．

5.1 今後の課題

本論文における，検索サイトの収集，フォーム情報の生成において主に以下のような問題点が浮かびあがった．

検索サイトの自動収集：本論文では Dnavi を用いて検索サイトの収集を行なった．ただし，収集方法は初めに “検索” というクエリを与え，得られた結果から 1 つずつ手動で調べるものだった．今後は，WWW 全体から検索サイトを自動的に収集する手法の開発を行なう必要がある．

様々なフォームへの対応：現在の DAISEn では，複数のテキストボックスがある場合は，それぞれを独立したテキストボックスとして扱う．また，ラジオボタンやプルダウンメニューなどテキストボックス以外のフォームには対応できない．今後は様々な検索サイトを扱うためにも，この様なフォームに対してもフォーム情報の作成を行なう必要がある．

謝辞 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．

統合検索システム DAISEn の開発にあたり，開発にたずさわった廣川研究室の卒業生の方々に感謝します．特に，中心的な役割を果たしていた古賀康則さん，酒井美由紀さんに感謝します．

文　献

[1] askOnce, http://www.askonce.com/
[2] BrightPlanet, The Deep Web: Surfacing Hidden Value, BrightPlanet White Paper, 2000.
[3] A. E. Howe and D. Dreilinger, Savvy Search: A Meta-Search Engine that Learns which Search Engines to Query, AI Magazine, Vol. 18, No. 2, pp, 19–25, 1997.
[4] P. Ipeirotis, L. Gravano and M. Sahami, PERSIVAL Demo: Categorizing Hidden-Web Resources, JCDL2001, 2001.
[5] P. Ipeirotis, L. Gravano and M. Sahami, Probe, Count, and Classify: Categorizing Hidden-Web Databases, ACM SIGMOD 2001, 2001.
[6] 国立国会図書館関西館データベース・ナビゲーション・サービス Dnavi, http://dnavi.ndl.go.jp/
[7] 古賀康則, 田口剛史, 廣川佐千男, 検索結果に含まれるタグパターン解析と抽出, 第 12 回データ工学ワークショップ, 2001.
[8] mamma, http://www.mamma.com/
[9] 中藤哲也, 酒井美由紀, 廣川佐千男, 検索サイトのための集合演算子の自動推定, 第 1 回情報科学技術フォーラム, 一般講演論文集第 2 分冊, pp. 9–10, 2002.
[10] T. Nakatoh, M. Sakai, Y. Koga and S. Hirokawa, Generation of Query URL for Search Sites, Proc. of *SSGRR* (CDROM), 2002.
[11] 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.
[12] 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, Directory Architecture for Integrated Search Engines, http://daisen.cc.kyushu-u.ac.jp/
[13] C. Sherman and G. Pric, The Invisible Web, Infomation Today, Inc., 2001.
[14] P. Pedley, The invisible web, ASLIB, 2001.
[15] T. Taguchi, Y. Koga and S. Hirokawa, Integration of Search Sites of the World Wide Web, Proc. of *the International Forum cum Conference on Information Technology and Communication*, Vol. 2, pp. 25–32, 2000.
[16] Vivisimo, http://vivisimo.com/
[17] 山田泰寛, 廣川佐千男, 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, 第 1 回情報科学技術フォーラム, 一般講演論文集第 2 分冊, pp. 11–12, 2002.