

複雑な検索機能を持つ検索サイトの動向調査

大森 敬介[†] 中藤 哲也^{††} 山田 泰寛[†] 原 由加里^{††} 廣川佐千男^{††}

[†]九州大学大学院システム情報科学府 〒812-8581 福岡市東区箱崎 6-10-1

^{††}九州大学情報基盤センター 〒812-8581 福岡市東区箱崎 6-10-1

E-mail: [†]{keisuke,yshiro}@matu.cc.kyushu-u.ac.jp, ^{††}{nakatoh,hara,hirokawa}@cc.kyushu-u.ac.jp

あらまし 近年, 単純なキーワード検索でなく, 複数の入力欄により複雑な検索ができる専門検索サイトが増えている. 国会図書館関西館のデータベース・ナビゲーション・サービス Dnavi には検索可能なデータベースが 2,800 件以上登録されている. 我々の研究室では, このような高度な検索サービスを統合するため, 入力項目の属性を自動的に推定するシステムの開発を行なっている. 本発表では, 統合検索についての研究の背景と, Dnavi に登録されている検索サイトについてのフォーム解析で得られた, 検索サイトの複雑化の動向について報告する.

キーワード メタデータ, メタサーチ, 検索エンジン, ディープウェブ

Trend Report of Search Sites with Complex Search

Keisuke OHMORI[†], Tetsuya NAKATOH^{††}, Yasuhiro YAMADA[†], Yukari HARA^{††}, and Sachio

HIROKAWA^{††}

[†] Department of Informatics, 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: [†]{keisuke,yshiro}@matu.cc.kyushu-u.ac.jp, ^{††}{nakatoh,hara,hirokawa}@cc.kyushu-u.ac.jp

Abstract General search engines provide a search facility to reach Web pages with a single query keyword. Besides these general search engines, there are many Web sites that provide Web interface to access their database with complex query, most of which are realized by multiple text inputs. Dnavi (Database Navigation Service of the National Diet Library) contains a list of 2,800 such searchable databases. Metasearch engines that integrate these specialized databases are good candidates for the search engines of next generation. One of the key issues in realizing metasearch engines is to discover schema of databases behind Web interface. This paper reports on the current situation of complex query forms and possibility to infer data schema by analyzing forms of HTML.

Key words Metadata, 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為に我々は, Yahoo!, Google 等の一般の検索エンジンを用いることが多い. しかしながら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必要な情報を持ったページが上位に表示されないなど, 検索結果の品質が大きな問題となっている.

一方, 企業などにおいては, 利用者へ直接情報を提供するため, 自サイト内の専門的な情報やデータベースについての検索サービスを提供するサイトが増えてきている. 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 **専門検索サイト**と呼ぶ. これらの専門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通常カバー出来ない. そのため, これらは Invisible Web [12], [13], Deep Web [1] あるいは Hidden Web [5], [6] と呼ばれる. CompletePlanet 社は, WWW 上で利用可能な 100,000 以上の探索可能なデータベースがあると推定している. これらの専門検索サイトは特定のテーマに限定した情報を提供しておりその品質は一般に高いため, 検索者の意図する分野に特化した専門検索サイトに対して統合検索を行なえば, 精度の高い情報を効率良く得られる.

WWW に存在する複数の検索エンジンもしくは専門検索サイ

トを統合するシステムは**メタサーチエンジン**と呼ばれ, Savvy-Search [4], mamma [22], vivisimo [25], askOnce [17] はその代表である。これらのメタサーチエンジンは複数の検索サイトを扱うため, 品質の良い結果が十分に得られるように思われるが, 技術的な問題から対象となる検索サイトは固定されている。

我々の開発している自動構築型メタサーチシステム **DAISEn**^(注1) (図1) は, 検索サイトを分析し, それらを統合するメタサーチエンジンを自動的に生成する [14], [18]。

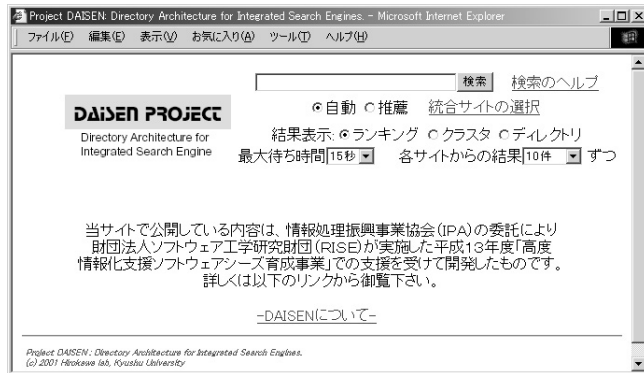


図1 自動構築型メタサーチシステム DAISEn

このシステムは, 次の4つの主要技術から成り立っている。

- クエリーパラメータの自動抽出技術 [11]
- クエリーの論理式の推定技術 [10]
- 情報抽出ラッパーの自動生成技術 [7]
- 検索サイトの特徴抽出技術 [3], [8]

最初の3つの技術が, 検索サイトのインタフェースの相違を隠蔽するラッパーの自動的な生成を可能にする。最後の1つの技術が検索サイトの自動的な特徴抽出である。生成されたラッパーを検索結果を囲む広告と飾りを削除するために用いることで, 我々はデータベースの純粋な特徴を得る。この特徴により適切な検索サイトを選択でき, より検索精度が高いメタサーチエンジンを生成することができる。

近年検索エンジンには, これまでとは異なる新しい方向性がみられるようになって来た。即ち複数の項目を用いた複雑な質問が行なわれ, そしてURLの単純なリストの代わりに, 幾つかの項目の集まりが返される。例えば, Amazon.com [16] は本のリストを返す, kakaku.com [21] はPCのリストと共にそれらの価格を返す, Travelocity [24] は指定されたエリアのホテルのリストを返す。それらの返す情報はURLの単純なリストではなく, いくつかの項目から構成された情報の集まりのリストだ。これらの専門的な検索エンジンの入力形式は, 一般的なサーチ・エンジンのものより極めて複雑である。それらはいくつかのキーワードを組み合わせて指定することを必要とし, それぞれのキーワードが異なった特質を表す。乗り換え検索の Jorudan [20] では, 出発と到着駅, 日時が必要である。また, ホテルのための検索エンジン Mytrip [23] では, チェックインの日付, チェックアウトの日付, 人数, 部屋数, 価格の上限と地域が必要である。

我々は, 現在の自動構築型メタサーチシステム DAISEn をより複雑な問い合わせ形式を持った検索サービスに対応させることを目指している。この目的の為に, 複雑な問い合わせ形式を持った検索エンジンを収集し, 調査している。我々は既に, 国立国会図書館のデータベースナビゲーションサービス **Dnavi** [19] 中に, 2,800以上の検索サイトがあると報告した [15], 更にそれらの検索サイトの構成について詳細な調査を行なった [9]。

本論文では, これらの専門検索サイトから典型的なものを対象として, 更に詳細な調査を行なう。また入力フィールドとその属性名の対応を自動的に抽出するツールを作成し, それを用いて検索統合の際に必要なメタデータの生成を行なう。

2. データベース・ナビゲーション・サービス Dnavi

国立国会図書館関西館の WebPage 中にあるデータベース・ナビゲーション・サービス Dnavi (図2) 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 における検索機能は本論文の目的であるメタサーチを提供するものではなく, 単純に目的とするデータベースへのリンク及びデータベースを見つけるための検索サービスである。Dnavi の収録データベース数は2003年6月10日現在約7,000件であり, このデータベース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。



図2 データベース・ナビゲーション・サービス “Dnavi”

既に我々は, Dnavi に掲載されているサイトから専門検索サイト 2,880 件を抽出し, その全体像を調査している [9], [15]。今回は, 複雑な質問形式を持ち特定の種類の検索サービスを提供する典型的な検索サイトとして図書検索サイトを対象を限定し, 詳細な調査を行なう。

Dnavi 中から抽出した 2,880 件の検索サイトから, 図書検索サイトのみを抽出したところ 938 件であった。抽出にあたっては全件に対して目視によるチェックを行ない確認した。

これらの中から, 図書検索サイトの典型的な一例を図3に示す。図書検索サイトの多くは, この例のように図書に関する複数の項目の1つあるいは複数を選択することで, データベース中の情報から指定にマッチする図書の一覧をユーザに提示する。

(注1) : <http://daisen.cc.kyushu-u.ac.jp/>

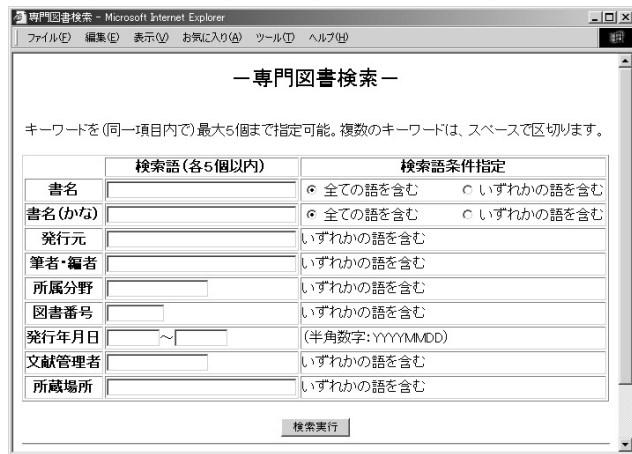


図3 図書検索サイトの例

3. 図書検索サイト

3.1 図書検索サイトの構成の傾向

抽出した 938 件の図書検索サイトについて、入力フィールドの数やその名前について詳細な調査を行なった。これは入力フィールドに関するメタデータの自動生成のための基礎情報を得るための調査である。本節では、この調査で得られた詳細なサイト情報について示す。

図書検索サイトにおけるテキスト入力フィールド^(注2)数を図4に示す。一見入力フィールドの少ないサイトも多いが、プルダウンメニュー^(注3)等を用いてフィールドの属性を切り替えるサイトが多く見受けられる。

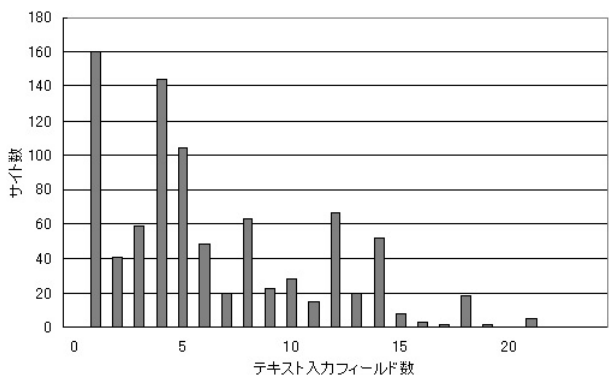


図4 図書検索サイトの持つテキスト入力フィールド数

プルダウンメニューの出現数で分類したサイト数を図5に示す。プルダウンメニューは検索条件等の設定の他に、テキスト入力フィールドの属性を指定する為にその直前に配置されることも多い。

ラジオボタン^(注4) (図6) は多くのサイトで検索条件の設定等に用いられていたが、セレクトメニューと同様にテキスト入力フィールドの属性を指定する為にも用いられていた。

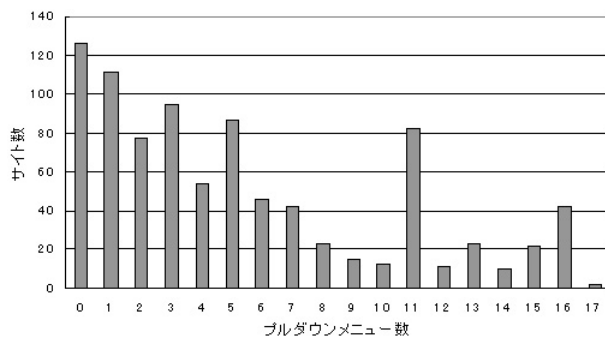


図5 図書検索サイトの持つプルダウンメニュー数

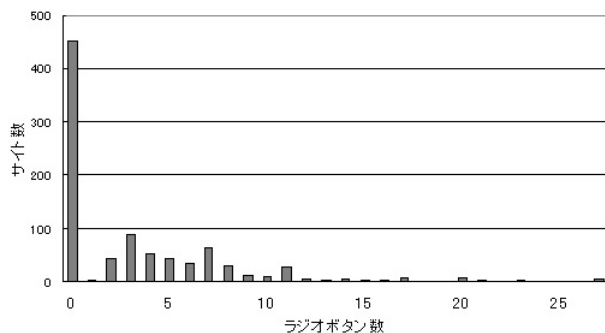


図6 図書検索サイトの持つラジオボタン数

チェックボックス^(注5) (図7) は多くのサイトで検索範囲の指定に用いられていた。

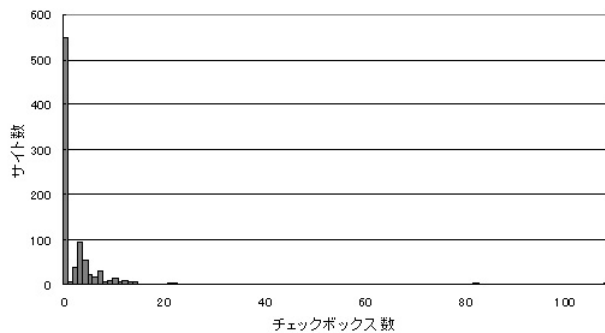


図7 図書検索サイトの持つチェックボックス数

3.2 図書検索サイトに現れる属性名及びメタデータ

今回入力要素の属性名を自動的に抽出するツールを作成した。それによって得られた属性名を手で分類し、図書検索に用いられる入力要素のメタデータを生成した。生成したメタデータ名とそれに属する属性名、更にはそれが出現するサイト数をまとめたものの一部を表1に示す。これらが単独で出現することは少なく、多くのサイトでこれらのいくつかの組合せとなっている。

(注2) : タグ <input type=text> で生成される要素

(注3) : タグ <select> で生成される要素

(注4) : タグ <input type=radio> で生成される要素

(注5) : タグ <input type=checkbox> で生成される要素

表1 図書検索におけるメタデータ

メタデータ名	属性名	サイト数
タイトル	ほんのなまえ, title, 図書名, フルタイトル, 本の名前, 本のなまえ, タイトル, 叢書名, 各巻書名, 書名 / title, 書名フルタイトル, 書名雑誌名, 書名タイトル, 書名 / タイトル, 書名叢書名, 書名, 書籍名, 書籍題名, ..	526
氏名	作者の名前, 雑誌記事著者名, 人名, 著者・編者, 編著者名, 編著者, 著者名, 著者名 (姓), 著者等 (姓), 著者 (姓), 書籍著者名, 姓, 名, ..	458
年月日	出版発行年, 和暦, 発行年, 制作年, 日付, 西暦, 刊行年, 所蔵時期, 出版年, ..	298
図書番号	図書館バーコード, 図書番号, 資料番号, 資料コード, issn/isbn, issn, isbn 番号, isbn, 雑誌コード, 分類番号, 分類コード, タイトルコード, 文書 id, 請求番号/資料 id, 請求番号, コード, 書誌番号, ..	360
条件名	ブラウズ条件, 検索項目間の条件, 検索条件, 資料区分絞り込み条件, 文庫絞り込み条件, 並び替え条件, 絞り込み検索条件, 項目間検索条件, 項目間の検索条件, 条件, ..	261

4. ま と め

本論文では、複雑な問い合わせ形式を持つ検索サイトとして図書検索サイトを取り上げ、その構成を詳細に調査した。更に入力要素の属性名を自動的に取り出すツールを開発し、それを用いて図書検索サイトから入力要素の属性名を抽出した。さらに、このデータを元に入力要素の属性名のメタデータを実際に構築した。

我々は、現在の自動構築型メタサーチシステム DAISEn をより複雑な問い合わせ形式を持った検索サービスに対応させることで、対応できる検索エンジンの範囲を広げるのみならず、より一般的な Web サービス全般に対する情報統合サービス [2] を提供することを目指している。その為には今後、より多くのタイプの専門検索エンジンを調査し、一般的な構造を明らかにすると共に、メタデータの自動生成手法を構築する必要がある。

謝 辞

本研究に関して多くの技術的援助を頂きました松永吉広氏、野口正人氏に感謝致します。また本研究の基礎となる DAISEn システムの開発に携わった多くの方に感謝致します。

文 献

- [1] BrightPlanet, The Deep Web: Surfacing Hidden Value, BrightPlanet White Paper, 2000.
- [2] Andreas Hess, Nicholas Kushmerick, Automatically attaching semantic metadata to Web services, Proc. IJCAI-03 Workshop on Information Integration on the Web, pp. 111-116, 2003.
- [3] S. Hirokawa, S. Watanabe, Y. Koga and T. Taguchi, Automatic Feature Extraction of Search Sites, Proc. SSGRR2001(CD-ROM).
- [4] 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.

- [5] P. Ipeirotis, L. Gravano and M. Sahami, PERSIVAL Demo: Categorizing Hidden-Web Resources, JCDL2001, 2001.
- [6] P. Ipeirotis, L. Gravano and M. Sahami, Probe, Count, and Classify: Categorizing Hidden-Web Databases, ACM SIGMOD 2001, 2001.
- [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] 中藤哲也, 古賀康則, 廣川佐千男, 検索統合のための検索サイト分類法, Proc. DBWeb2001, pp. 225-228, 2001.
- [9] T. Nakatoh, K. Ohmori, Y. Yamada and S. Hirokawa, COMPLEX QUERY AND METADATA, Proc. ISEE2003, pp. 291-294, 2003.
- [10] 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 (FIT2002), 一般講演論文集第 2 分冊, pp. 9-10, 2002.
- [11] T. Nakatoh, M. Sakai, Y. Koga and S. Hirokawa, Generation of Query URL for Search Sites, Proc. of SSGRR2002w (CDROM), 2002.
- [12] P. Pedley, The invisible web, ASLIB, 2001.
- [13] C. Sherman and G. Pric, The Invisible Web, Information Today, Inc., Medford, New Jersey, 2001.
- [14] 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.
- [15] 山田泰寛, 松永吉広, 野口正人, 中藤哲也, 廣川佐千男, 統合検索システム DAISEn での検索サイトフォーム分析, 情報処理学会研究報告 2003-DBS-131(II)(77)(夏のデータベースワークショップ DEWS2003), pp.311-318, 2003.
- [16] Amazon.com, <http://www.amazon.com/>
- [17] askOnce, <http://www.askonce.com/>
- [18] 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/>
- [19] 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/>
- [20] Jorudan, <http://www.jorudan.co.jp/>
- [21] kakaku.com, <http://www.kakaku.com/>
- [22] mamma, <http://www.mamma.com/>
- [23] Mytrip, <http://www.mytrip.net/>
- [24] Travelocity, <http://www.travelocity.com/>
- [25] Vivisimo, <http://vivisimo.com/>