WebDBの組み合わせによる情報収集ツールの構築を目指して ： 山下記念研究賞受賞の報告と研究の紹介

中藤, 哲也
九州大学情報基盤研究開発センター

大森, 敬介

廣川, 佐千男

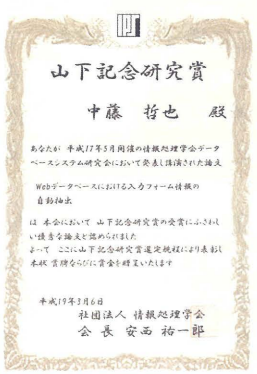
https://doi.org/10.15017/1470723

出版情報：九州大学情報統括本部ITマガジン. 1 (3), pp.52-58, 2008-03. 九州大学情報統括本部
バージョン：
権利関係：

WebDBの組み合わせによる情報収集ツールの構築を目指して
～山下記念研究賞受賞の報告と研究の紹介～

中藤 哲也 *　　大森 敬介　　廣川 佐千男

我々の研究グループでは，検索機能を持ったWeb上のデータベース（Webデータベース）の機能を組み合わせて，自由に情報検索ツールを組み立てるための技術に関する研究を行っております．その研究の一環として，平成17年5月開催の情報処理学会データベースシステム研究会にて発表した「Webデータベースにおける入力フォーム情報の自動抽出」において，情報処理学会平成18年度山下記念研究賞を頂くことができました．本稿では，その研究の背景と概要について紹介いたします．また，本研究で構築するツールのプロトタイプとして作成した国内の情報処理系学会の論文情報抽出ツールを公開しておりますので，それに関して紹介したいと思います．


山下記念研究賞

中藤　哲也　殿

あなたが 平成17年5月開催の情報処理学会データベースシステム研究会において発表し推薦された論文

Webデータベースにおける入力フォーム情報の自動抽出

は 本会において 山下記念研究賞の受賞にふさわしい優秀な論文と認められました よって ここに山下記念研究賞選定規程により表彰し 本状 賞牌ならびに賞金を贈呈いたします

平成19年3月6日
社団法人 情報処理学会
会 長 安西 祐一郎

1　山下記念研究賞について

山下記念研究賞とは，情報処理学会の研究会および研究会主催シンポジウムにおける研究発表のうちから特に優秀な論文を選び，その発表者に贈られるものです．受賞者は該当論文の登壇発表者である会員で，年齢制限はありません．本賞の選考は，表彰規程，山下記念研究賞受賞候補者選定手続および山下記念研究賞推薦内規に基づき，各領域委員会が選定委員会となって行います．

平成18年度は33研究会の主査から推薦された計57編の優れた論文に対し，慎重な審議を行い，決定されたうえで，第519回理事会（平成18年7月）および調査研究運営委員会に報告されたものです．平成18年度の受賞者は57名で，平成19年3月6日に早稲田大学で開催された第69回全国大会の席上で表彰状，賞牌，賞金が授与されました．

我々の研究は次の理由で推薦して頂きました．

「受賞対象論文は，Webブラウザに表示される入力フォームから検索用キーワードを指定して利用するWebデータベース(検索サイト)の入力属性を自動的に抽出する手法を提案し，様々な実際のWebデータベースを対象として有効性を検証している．提案手法によりWebアプリケーションからWebデータベースを利用することが大変容易になると考えられる．実用化にはさらに抽出精度を向上する必要があるが，有用性は高く今後の進展が望まれる．よって本論文発表者を山下記念研究賞受賞候補者として推薦する．」

* 九州大学 情報基盤研究開発センター, Email: nakatoh@cc.kyushu-u.ac.jp

2 研究の背景と概要

Google などの一般検索サイトや特定の情報に特化した専門検索サイトを使って，何か特定のテーマについての調査を行う場合には，一つのキーワードだけを使った一度だけの検索で作業が終ることは余りありません．例えば，ある地域のレストランリストを検索して，次に各レストランのメニューや価格に関する情報を集めたり，販売中の中古車の一覧検索から幾つかの車の詳細情報を集めて比較するような場合が多い筈です．表示された何十件もの検索結果を見ることで，重要な人名や関連するキーワードを読み取り，より広い検索やより細かな検索を行います．適切なメモを残しながらこのような操作を繰り返して，納得できる検索結果のリストを作ります．それらの検索は，ある一つの検索サイトで続けて行う事もあれば，別の検索サイトに移って行う事もあります．

このように検索エンジンを反復的に利用する場合，一般の検索エンジンの利用と専門検索エンジンの利用では状況が大きく異なります．一般の検索エンジンでは，検索結果は多様な Web ページであり，再検索のための新たなキーワードを獲得する一般的な手立てはありません．一方，専門検索サイトの検索結果は多くの場合，そのサイトの背後にあるデータベースから得られた同質なデータです．例えば，文献検索のサイトでは，人名やキーワードを与えて得られる結果は，単なる Web のページではなく，著者，タイトル，雑誌名，ページ，出版年などの項目からなる文献データです．我々が網羅的に文献検索を行う場合には，1 回目の検索結果が得られてもそれで終りでなく，そこで得られる情報をもとに更に検索を続けることが多いと言えるでしょう．一つの論文を見つけると

- 著者や共著者が他にどのような論文を書いているか
- その論文はどのような論文を引用しているか
- その論文がその後，どのように引用されているか
- 関連研究で重要なキーワードはなにか
- 著者らのホームページはどこか
- 関連するプロジェクトがあるか

などを繰り返し調べます．つまり，専門検索サイトを使って反復的に検索を行う時には，文献データという構造情報から著者やタイトルという部分的情報を抽出して利用しているわけです．例えば，DBLP [*1]や CiteSeer [*2]のような文献検索のサイトでは，文献リストを検索結果として返すだけでなく，このようなユーザーの操作を先取りし，著者ごとに分類したページやそのようなページを動的に生成する URL へのリンクが提示されているので，効率よく関連研究の調査を行うことができます．

一つの専門検索サイトを反復的に利用できるのは，出力情報データの属性に入力として使える項目があるからです．複数の専門検索サイトの統合検索（メタサーチ）が考えられるのは，それらの入力データと出力データの構造が類似しているからです．また，ある検索サイトの出力データの属性として人名が含まれれば，人名を検索キーワードとする他の検索サイトの入力と結合して利用することを考える事ができます．

この後の章では，このように専門検索サイトを入力データ構造と出力データ構造で規定される抽象的部品として捉え，それらを結合することにより新たな検索ツールを構成する方法を紹介します．また，その為に必要

[*1] http://dblp.uni-trier.de/
[*2] http://citeseer.ist.psu.edu/

な技術の一つとして，今回の受賞内容となった入力フォーム情報の自動抽出に関する技術を紹介します．最後に，本方式によるアプリケーションの実例の一つとして，情報処理学会，電子情報通信学会，人工知能学会，及び日本ソフトウェア科学会の各学会の論文検索システムを対象とするプロトタイプを実装しましたので，その紹介を致します．

3　専門検索サイトの部品化とその効果

専門検索サイトは一般に，内部のデータベースが持つ情報を扱っています．そのため，データベースの持つフィールドの一部を入力として受け取り，それを含んだレコードのリストを出力するものが多い傾向にあります．例えば，図 1 に示す図書検索サイトでは，入力フィールドとして書籍に関する複数のフィールドがあり，それらの一部を指定する事で，一致する書籍データの一覧をユーザに提示します（図 2）．

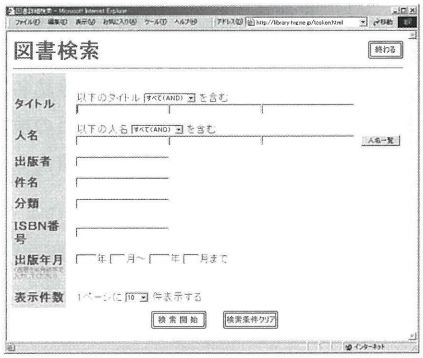


図 1　専門検索サイトの例

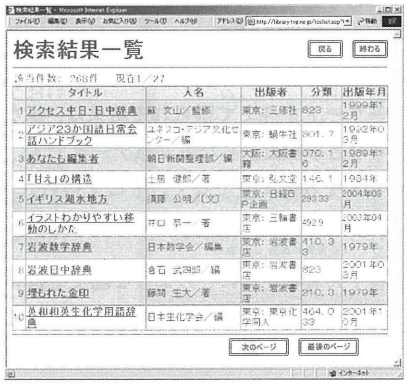


図 2　専門検索サイトの検索結果の例

検索サイトの多くはこの例にあるように，入力項目と出力項目をペアとすることでその機能を表現できます（図 3）．更に各検索サイト専用のラッパー [*3] を組み合わせる事により，他と組み合わせ可能な部品としての基本機能を持たせることができます（図 4）．我々はこれまでに 2,880 件の専門検索サイトを収集していますが [4]，それらのうち，幾つかの検索サイトについての入出力項目を，実例として表 1 にまとめました．

図書検索サイト

入力項目	出力項目
タイトル	タイトル
人名	人名
出版者	出版者
件名	分類
分類	出版年月
ISBN番号	
出版年月	

図 3　検索サイト機能の模式図

検索サイトを入力と出力の組として捉えると，(1) 入力の統合，(2) 出力の統合，(3) 入力と出力の結合，の 3 通りの組み合わせ方法を考えることができます．従来のメタサーチエンジンは (1) の入力の統合だけを実装

[*3] 詳細を隠蔽して，必要とされる機能だけを外部に提供するツールの総称．
ここでは検索サイトによって異なる機能や入出力方法を隠蔽し，統一されたインターフェースを用いて機能を提供するプログラム．

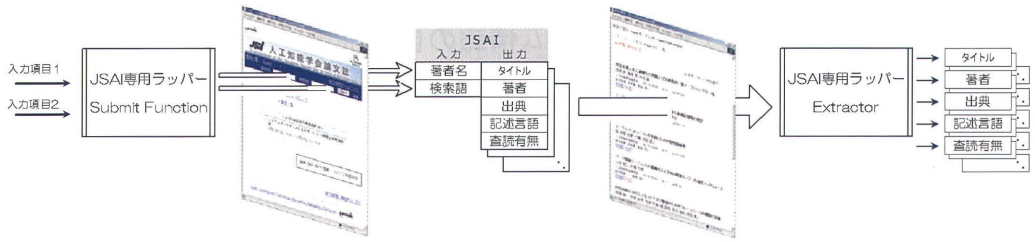


図 4　検索サイトの部品化

表 1　専門検索サイトの入出力の例

サービス名	入力項目	出力項目
図書検索	タイトル，著者， 出版者	タイトル，著者，出版者， ISBN，発行年
施設検索	施設名	施設名称，施設分類 住所，TEL，管理
塾 予備校検索	塾・予備校名， 駅名，出版者	学校名，教室名 指導形態，対象，沿線駅名
野菜生理障害 事例検索	キーワード	病名
特許検索	検索語句	出願番号，発明の名称
最寄りの お風呂屋さん	最寄り駅	浴場名，路線・最寄り駅名，道順， 住所，営業時間，定休日，特徴，URL

したもので，各々の検索結果は単純に（あるいはランキング付で）リスト化されているだけです．(2) の出力の統合では出力結果の各フィールドの意味（メタデータ）を使って，例えば価格の比較が可能な，一覧表として検索結果が表示できます．(3) の例としては，求人情報検索で得られた企業について，その企業の業績や動向を株価の検索で調べるような場合があります．

また検索サイトの部品化により，部品の結合に関して一般的プログラミングを考える事が可能となります．例えば，「二つの図書館に対する検索を実現する」ためのスクリプトを書くことを考えます．従来提案されているシステム，例えば MetaCommander [2] では図書館 A，図書館 B それぞれについて，CGI にどのようにパラメータを渡すか，出力の HTML から本の情報をどのように抽出するかを，プログラム中に直接埋め込む必要があります．一方，本稿で提案する方式ではこの問題を，

(a) 図書館 A，図書館 B を入出力データ構造の組として捉える
(b) 二つの図書館情報検索機能の結合方式を記述する
(c) 統合したシステムの入出力のインターフェースをマッチングさせる

という 3 つの部分に分離して解決します．(a) のためには，それぞれの図書検索サイトについて個別にラッパーを構築する必要があります．しかし，対象とする図書館が変わったとして，(b), (c) の部分は変更する必要はありません．(b), (c) の部分は「複数の Web サービスをどのように組み合わせるか」という一般的なプログラムとして，より抽象的に構成することができます．

4　Web データベースの入力フォーム情報の自動抽出

本節では，山下記念研究賞を頂いた研究に関して概略を紹介致します．

専門検索サイトを部品化し，それらを連携して利用するには，主に（1）WebDB への検索実行の自動化，（2）WebDB の検索結果の出力ページからのレコードの抽出 *4，の 2 つの機能が必要です．

従来型のデータベース，あるいは Web サービスであれば，データの受け渡し方法，及びデータの構造（データスキーマ）が明示的に与えられています．しかし，WebDB ではブラウザ経由での人間による利用しか想定されていないため，それらの情報は一般に明示されません．したがって，データの受け渡し方法，及びメタデータは，HTML を用いて記述された入力インタフェースの画面から自動的に抽出する必要があります．

受賞の対象となった研究ではこれらのうち，WebDB への検索実行の自動化に必要な機能として，入力フォームの自動的な分析，及び入力フィールドの項目名（メタデータ）抽出を扱いました．

まず，属性の異なる複数の入力項目を持った入力フォームの現状についての調査を行いました．従来の研究 [6, 1, 8] においては，入力フィールドのメタデータとして，各入力フィールドに近接する文字列を想定しています．しかしながら，これまでの調査 [9] により，複数の入力フィールドを持つサイトでは，多くの例で TABLE タグを用いて表示画面の構造を記述している事が分かりました．具体的には，調査した 2,800 サイト中 1,359 サイト，すなわち 48.5% のサイトにおいて，入力フィールドを内部に持つ TABLE タグが存在していました．また一般的な HTML の表からの情報抽出に関する研究 [3, 7] などからも，同様の事が示唆されます．従って，本研究では，WebDB の入力フォームの多くが表で記述されている事実に基づいて，新たに考案した抽出方法を示しました．その上で，国内の 2,800 件の WebDB から無作為に選んだ 134 件のサイトに対し人手により抽出したメタデータと提案手法で抽出したメタデータとの比較，及びナイーブな手法と提案手法との比較を行い，提案手法の抽出性能を評価しました．

本研究に関して，より詳しい内容をお知りになりたい方は，研究会予稿集 [10]，あるいは論文 [11] を参照して下さい．

5　論文検索システム（プロトタイプ）の紹介

検索サイトを部品として，それらを組み合わせた情報収集ツールのプロトタイプを作成しました．ここでは，その情報収集ツールを紹介します．本ツールは，次に示す国内の情報処理系の各学会の論文検索サイトを対象にして，論文の情報を収集する事を目的としています．

- 情報処理学会電子図書館*5
- 電子情報通信学会 和文論文誌*6
- 電子情報通信学会 英文論文誌*7
- 人工知能学会論文誌*8

*4 この点に関しての研究も行っています [5].
*5 http://www.bookpark.ne.jp/ipsj/, 会誌，英文誌，研究報告，論文誌 (ジャーナル及びトランザクション)，欧文誌を含む
*6 http://search.ieice.org/jpn/search-j.html
*7 http://search.ieice.org/search.html
*8 http://tjsai.jstage.jst.go.jp/ja/

- 日本ソフトウェア科学会 J-STAGE[*9]

本システムは主に三つの機能から成り立っています．それらは，(1) 複数の検索サイトに対して同時に検索を行い，結果を統合してユーザに提示する機能，(2) 結果中の著者名を抽出し，リスティングする機能，(3) リスティングされた著者名をキーとした次のステップの検索を提供する機能，です．

(1) は，いわゆるメタサーチの機能と同じです．個々の検索サイトに対するラッパーにより入出力の違いを吸収し，得られた複数の結果を組み合わせてユーザに提示します．(2) は，出力結果のページの解析により著者名及び共著者名を抽出し，それらを一覧表としてユーザに提示します．これには (3) の機能へのリンクも持たせています．(3) は，得られた情報を元に繰り返し検索を行う機能です．得られた論文一覧中の著者名をクリックする事で再び新たな検索を行い，その著者に関する論文情報を提示します．

この 3 つの機能のうち，(2), (3) は文献検索システム DBLP で用いられているものと同じです．検索結果に対するこのような処理を含む機能は，利用している DB の直接的アクセスが必要なので，DBLP のように通常システム中に組み込まれなければなりません．一方，我々の提案する方式では，独立した文献検索システムを統合するだけでなく，この (2), (3) の機能をそれぞれのシステムの外部に構成することができます．

これらの機能のデータ結合の模式図を図 5 に示します．

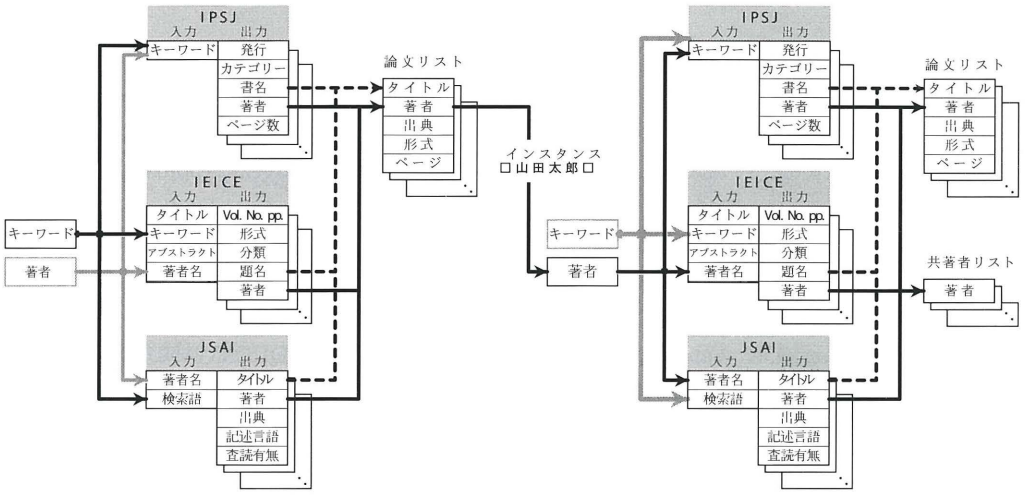


図 5　データ結合

本システムの基本動作を見る事にします．先ず最初に，著者名による検索か，キーワードによる全文検索を行います（図 6）．本システムは入力された条件（キーワード or 著者名）を各検索サイトの要求する入力フォームに合わせて変換します．そのフォームを各検索サイトへ送り，それぞれ検索を行います．得られた結果は，各検索サイト毎のラッパーでフィールド単位に分割し，全てのサイトからの結果を一つの表にまとめてからユーザに提示し，同時に次の検索へデータを渡すためのリンクを生成し，各著者名に関連付けます（図 7）．

[*9] http://www.jstage.jst.go.jp/browse/jssst/-char/ja/

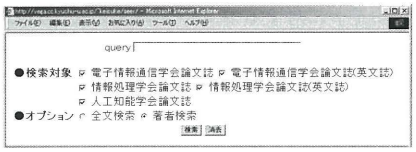


図 6　著者検索システム（プロトタイプ）

図 7　検索結果の例

ユーザは，参照したい著者名をクリックするだけで，順次関連情報を検索して行く事ができるようになります．我々は本システムを，`http://matu.cc.kyushu-u.ac.jp/guruguru` にて公開していますので，よかったらご利用下さい．

参考文献

[1] He, H., Meng, W., Yu, C. and Wu, Z., "Automatic Integration of Web Search Interfaces with WISE-Integrator", *VLDB Journal*, Vol.13, No.3, pp.256-273, 2004.

[2] Kitamura, Y., Noda, T. and Tatsumi, S., "Single-agent and Multi-agent Approaches to WWW Information Integration," *Multiagent Platforms, Lecture Notes in Artificial Intelligence*, Vol. 1599, Berlin et al.: Springer-Verlag, 133-147, 1999.

[3] Lerman, K., Knoblock, C. and Minton, S., "Automatic Data Extraction from Lists and Tables in Web Sources", *Proc. of ATEM-10, IJCAI*, 2001.

[4] Nakatoh, T., Ohmori, K., Yamada, Y. and Hirokawa, S., "COMPLEX QUERY AND METADATA," *Proc. of ISEE2003*, pp. 291-294, 2003.

[5] Nakatoh, T., Yamada, Y. and Hirokawa, S.: "Automatic Generation of Deep Web Wrappers based on Discovery of Repetition", *Proc. of AIRS2004*, pp.269-272, 2004.

[6] Raghavan, S. and Garcia-Molina, H., "Crawling the HiddenWeb", *Proc. of the 29th International Conference on VLDB*, pp.129-138, 2001.

[7] Yoshida, M., Torisawa, K. and Tsujii, J.: "Integrating Tables on the World Wide Web", *Trans. of the JSAI.* Vol.19, No.6, pp.548-560, 2004.

[8] Zhang, Z., He, B. and Chang, K. C., "Understanding Web Query Interfaces: BestEffort Parsing with Hidden Syntax", *SIGMOD2004*.

[9] 大森敬介, 中藤哲也, 山田泰寛, 原由加里, 廣川佐千男, "複雑な検索機能を持つ検索サイトの動向調査", *Proc. of DEWS2004*, I-1-05, 2004.

[10] 中藤哲也, 大森敬介, 廣川佐千男, "Web データベースにおける入力フォーム情報の自動抽出," *IPSJ SIG Technical Reports*, 2005-DBS-136. pp.87-94. 2005.

[11] 中藤哲也, 大森敬介, 廣川佐千男, "WebDB の QueryForm におけるメタデータ自動抽出," *DBSJ Letters* Vol.5, No.2, pp.97-100, 2006.